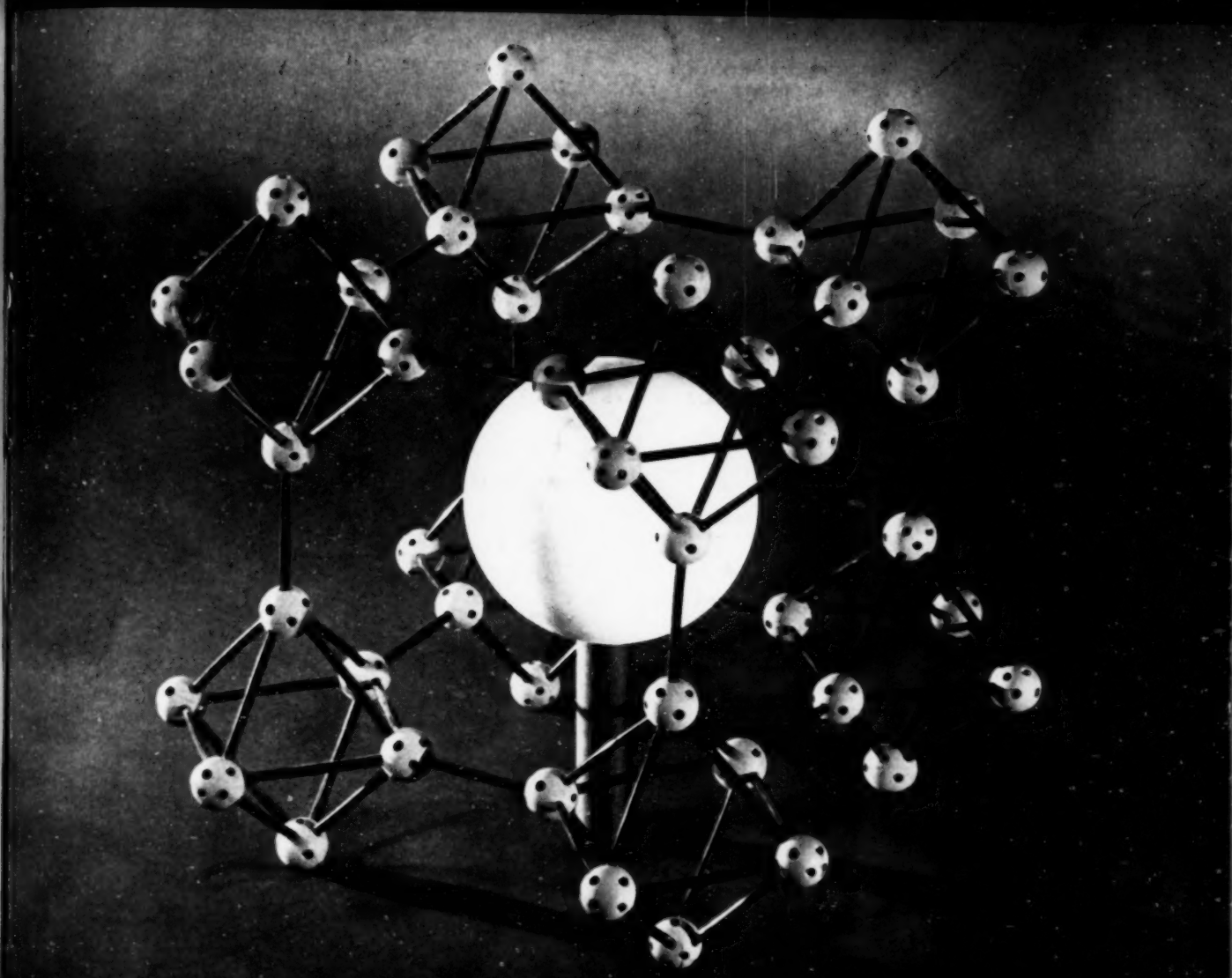
JOURNAL OF APPLIED PHYSICS



JOURNAL OF APPLIED PHYSICS

Vol. 22, No. 3 MARCH, 1951

In This Issue

Page

Effects of Radiation on Materials J. C. Slater 237

Contributed Original Research

Electron Optical Exploration of Space Charge in a Cut-Off Magnetron D. L. Reverdin 257
Note on Reflection and Transmission Bela A. Lengyel 263
Reflection and Transmission at the Surface of Metal-Plate Media Bela A. Lengyel 265
Experimental Determination of the Reflection Coefficient of Metal-Plate Media J. Ruze and M. Young 277
On Amplitude Bounds for Certain Relaxation Oscillations Nelson Wax 278
Theory of the Envelope Type of Thermal Conductivity Tests Arthur L. Loeb 282
Long-Range Order in Beta-Brass and Cu_2Au D. T. Keating and B. E. Warren 286
Phase Equilibria in an Ordering Alloy System J. B. Newkirk, R. Smoluchowski, A. H. Geisler, and D. L. Martin 290
Boride Cathodes J. M. Lafferty 299
Transmission-Line Equivalent of Electronic Traveling-Wave Systems W. E. Mathews 310
Thickness-Shear and Flexural Vibrations of Crystal Plates Raymond D. Mindlin 316
Correlating Modulus for Fluid Resistance in Accelerated Motion H. W. Iversen and R. Balent 324
Cut-Off Frequency in Two-Dielectric Layered Rectangular Wave Guides J. Van Bladel and Thomas J. Higgins 329
Maxwell's Demon Cannot Operate: Information and Entropy. I L. Brillouin 334
Physical Entropy and Information. II L. Brillouin 338
Metallographic Study of Diffusion Interfaces G. C. Kuczynski and B. H. Alexander 344
Scattering and Absorption of Gamma-Rays M. S. Plesset and S. T. Cohen 350

Letters to the Editor:

Elastic Parameters of Beta-Brass Robert A. Artman and Donald O. Thompson 358
Barium Oxide on Tungsten Cathode Interface H. P. Rooksby and E. G. Steward 358
Barium Oxide on Tungsten Cathode Interface Eugene B. Hensley and John H. Affleck 359
Water Movement Induced by Compaction of Clastic Sediments Robert A. Howard and Carl E. Bleil 360
Impact Flash at High Speed Water Entry J. H. McMillen, R. L. Kramer, and D. E. Allmand 360
Crystal Structure of the Sigma-Phase in the Co-Cr System J. S. Kasper, B. F. Decker, and J. R. Belanger 361
Cathodoluminescence of Zinc Orthosilicate with Manganese Activator S. Larach and R. E. Shrader 362
Investigations on the Capturing of Electrons in a 30-Mev Betatron R. Wideröe 362
Thermal Expansion Coefficients and the Melting Point of Cubic Elements D. van der Reyden 363
On the Semiconductor Bornite C. W. Horton 364
Epitomization of the Basic Theory of the Generalized Schwarz-Christoffel Transformations as Used in Applied Physics Thomas James Higgins 365

Index to Advertisers xxx

BOARD OF EDITORS

Elmer Hutchisson, *Editor*

Earle C. Gregg, *Assistant Editor*

Associate Editors

Howard H. Aiken
Lyle D. Borst
James Hillier
Leonard B. Loeb
W. James Lyons
Eric Reissner
Raymond J. Seeger
Frederick Seitz
Charles S. Smith

Publication of the

American Institute of Physics

OFFICERS

George R. Harrison, *Chairman*
George B. Pegram, *Treasurer*

GOVERNING BOARD

J. W. Beams
C. Paul Boner
Wallace R. Brode
W. F. Fair, Jr.
S. A. Goudsmit
G. P. Harnwell
George R. Harrison
F. V. Hunt
Rudolph Kingslake
Paul Kirkpatrick
F. W. Loomis
J. R. Oppenheimer
George B. Pegram
I. I. Rabi
Duane Roller
Harold K. Schilling
Frederick Seitz
Francis G. Slack
John C. Steinberg

Consultant on Printing

Melvin Loos

ADMINISTRATIVE STAFF

Henry A. Barton, *Director*
Wallace Waterfall, *Executive Secretary*
Theodore Vorburger, *Advertising Manager and Assistant Treasurer*
Ruth F. Bryans, *Publication Manager*
Betty H. Goodfriend, *Assistant Secretary*
Eileen B. Neuberger, *Circulation Manager*
Kathryn Setze *Senior Accountant*

The Journal of Applied Physics, published monthly at Prince and Lemon Sts., Lancaster, Pa., is devoted to physics in its role as the science basic to other natural sciences and to the arts and industries. Previous to 1937 this publication was known as Physics. The Journal of Applied Physics publishes editorials and reviews, as well as technical papers of applied physics. Articles appearing in it are indexed regularly, according to their field, by Physics Abstracts, Chemical Abstracts, Engineering Index, Inc., Industrial Arts Index, etc.

Manuscripts should be submitted to Elmer Hutchisson, Editor, Case Institute of Technology, Cleveland 6, Ohio. Unless otherwise stated, submission of a manuscript is a representation that it has been neither copyrighted, published, nor submitted for publication elsewhere. The authors' institutions are requested to pay a **publication charge** of $8 per page which, if honored, entitles them to 100 reprints without further charge. Instructions will be sent with galley proofs.

Subscriptions, renewals, and orders for back numbers should be addressed to the American Institute of Physics, 57 East 55 Street, New York 22, New York.

Subscription Price	U.S. and Canada	Elsewhere
To members of the American Institute of Physics and Affiliated Societies	$10.00	$11.00
To all others	12.00	13.00

Back Number Prices

Yearly back number rate when complete year is available: $10.00. Single copies: $1.00 each.

Changes of address should be sent to the American Institute of Physics; **corrected proofs** should be sent to the Publication Manager, American Institute of Physics. **Advertising rates** supplied on request. Orders, advertising copy, and cuts should be sent to the American Institute of Physics.

Entered as second class matter January 22, 1937, at the Post Office at Lancaster, Pennsylvania, under the Act of March 3, 1879. Accepted for mailing at the special rate of postage provided for in paragraph (d-2), section 34.40, P. L. & R. of 1948, authorized May 2, 1932.

Journal
of
Applied Physics

Volume 22, Number 3 March, 1951

The Effects of Radiation on Materials

J. C. Slater

Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts

HIGH energy radiation, whether in the form of charged ions, electrons, neutrons, or electromagnetic radiation, can have permanent damaging effects on materials which are of practical importance in the design of nuclear reactors. This paper is a survey of the types of damage produced by various kinds of radiation in various solids, with the relations to the theory of solid-state structure, and a discussion of the types of experiments which have been or could be performed to investigate radiation damage. It covers the work of many persons, many of whom have contributed to the classified literature. It is an excerpt from a report prepared by a group consisting of B. L. Averbach, D. S. Billington, J. W. Irvine, Jr., W. E. Johnson, A. R. Kaufman, A. W. Lawson, Jr., J. R. Low, S. Untermyer, and the author, which carried out a survey of the subject for the Atomic Energy Commission during the summer of 1949. This excerpt, which was written by the author, has been declassified by the Atomic Energy Commission. The author wishes to thank the other members of the group, and numerous members of the staffs of the various AEC laboratories, for valuable discussions and advice.

Radiation damage is the result of displacement of atoms in a solid by particles passing through it, of the particles which remain embedded in the solid as impurities, and of the ionization produced by the particles. To understand it, we must know what effects are produced in solids by the atomic displacements and the impurities, and these effects are similar whether they are produced by radiation or by other mechanisms. We shall therefore start by a short review of the current state of knowledge of solid-state behavior, giving particular weight to displacements and impurities. When we have looked into the theories of these things, and the experimental methods of finding out about them, we shall be able to specialize to the case of damage produced by irradiation.

The fundamental fact about a solid is that under ordinary circumstances its behavior is a unique function of the positions of its atoms, no matter how its atoms came to those positions. A solid is made of two sorts of constituents: the nuclei of the atoms, and the electrons. With given positions of the nuclei, the electrons are capable of a great many stationary states; but their transitions from one state to another are ordinarily so fast that they will have attained a configuration of thermal equilibrium in times short compared to the times concerned in our experiments, which is the basis of our statement that the behavior of the solid is uniquely determined by the positions of the atoms (that is, of their nuclei). The only exceptions are, first, the case where we are dealing with extremely short periods of time, as the time immediately after the passage of a fast particle through matter, in which case the electrons may not have had time to come to equilibrium, even in spite of their very rapid approach to equilibrium; or secondly, the case of a very good insulator, in which electrical conduction by electrons is so slow that non-equilibrium charge distributions, as for instance the continued existence of space charges not corresponding to thermal equilibrium, may persist for measurable times. We shall make specific mention of these cases when they come up. Aside from them, however, our generalization remains true.

It is clear, then, that any study of a damaged solid can be profitably broken up into two parts: how did the atoms get where they are, and what effect do the atomic positions have on the properties of the material? The meeting ground of the two questions is the straightforward problem, where are the atoms? Of course, in contrast to the electrons, we cannot assume that the atoms are in positions of thermal equilibrium, for many processes of atomic motion are so slow, at low temperatures, that the atoms can remain in unstable, or metastable, positions for a very long time. By far the most

straightforward way to find where the atoms are is by looking at them, that is, by use of x-ray, electron, and neutron diffraction. These methods cannot tell us as much as our eyes can about really visible things, but we shall indicate in a later section that they can go rather far in determining the nature of displacements in solids. All other forms of information are more indirect; most of them are in the nature of observations of various physical properties affected by the displacements, and their interpretation depends on how good our theories of those physical properties may be.

We shall now proceed with our review of solid-state behavior, answering in general the question, given the atomic positions, how does the material act? We shall then go back to the other question, how did the atoms get where they are, by processes of irradiation?

I. THE BEHAVIOR OF AN IMPERFECTLY CRYSTALLINE SOLID

The characteristic properties we are interested in depend on crystal imperfections, but we can best start by asking about a perfect crystal. Here we meet two methods of discussing it, the atomic picture and the band picture, two alternative theoretical approaches which converge to similar results, but each useful in its own sphere.

1. The Atomic and Band Pictures of Solids

One picture of solids, the atomic picture, is particularly useful for discussing their mechanical properties. In it, we assume each atom acting on its neighbors by forces of various sorts: electrostatic forces, if they happen to be charged ions; valence attractions, if they have shared electrons between them; electronic repulsions, as they are brought too closely into contact, and the electron clouds surrounding each atom begin to get in each others' way; van der Waals attractions, weak attractions effective between any pair of atoms at large distances of separation. All these types of forces are found between the two atoms in a diatomic molecule, or between two atoms colliding in a gas. We cannot, however, treat the solid just as if we superposed the forces acting between each pair of atoms, for the presence of some atoms bound to a given atom will modify the capacity of that atom to exert forces on another neighbor. The phenomenon of saturation of valence is the most familiar example of this action: the way in which, when an atom has made a number of bonds with its neighbors equal in number to its valence, it cannot make additional bonds with other neighbors. If the valences of an atom are saturated, its electronic structure is in a sense frozen into position, not capable of easy change; whereas if it has unsaturated valences, it is easy for other bonds to form, and there is a certain flexibility in its structure.

In a crystal of a valence compound, the valence bonds between the atoms of the molecules are usually saturated, so that their electronic structure is rather rigid, and electrons cannot flow from one part of a molecule to another, so that they do not conduct electricity. Molecules are held to each other by the weak van der Waals forces, so that the materials as a whole are insulators, and they melt and vaporize into molecules as the temperature is raised. In a metal, on the other hand, the valences bonding one atom to another are unsaturated, no molecules form, and the electrons move rather freely from one valence bond to another, or from one part of the crystal to another, so that the material is a conductor of electricity. In an ionic crystal, such as NaCl, or the oxides like BeO or BaO, the crystal is made up of ions forming closed electronic shells; these ions are held together by electrostatic attraction, kept from collapsing by the electronic repulsions, and there are no valence bonds, no flexibility of the electronic structure, and no electronic conductivity.

In considering the mechanical properties of a solid, this atomic picture is a valuable one. In the simplest type of distortion, such as a homogeneous compression, the atoms are all forced closer together, their repulsive forces become stronger than their attractions, and they exert a pressure. In a homogeneous elastic shear planes of atoms tend to slip over each other, and by considering the variation of the forces with distance we can find the resistance to shear. In more extreme displacements, such as we have when an individual atom gets pushed into an interstitial position or pulled out of a lattice, we can still use the same type of argument; but in this and other cases we must be careful to modify the simple interatomic forces to take account of saturation of valence. Thus, it is clear that if an atom is a crystal finds itself near just enough neighbors to saturate its valences, the approach of an additional atom will have quite different results from what it would have done if there had been fewer neighbors.

For the electrical properties, we have seen that we get a correct general view of the conductivity of substances from the atomic picture. We can also interpret the magnetic behavior, to some extent: the inner electronic shells of some elements are not filled, they have a net magnetic moment; and since the shells are not filled, they have a certain flexibility, showing itself in the possibility that the magnetic moment may orient itself in different directions, producing magnetic effects. On the other hand, there are some electrical properties which can hardly be explained with any ease by this atomic picture. As a typical example we may mention the three materials C (diamond), Si, Ge. All three materials crystallize in the same diamond structure, which seems to be a typical saturated valence structure, each of the atoms forming four well-defined valence bonds to its four adjacent neighbors, just saturating its valence. We should certainly expect these materials to be insulators, according to previous indications, and diamond is, but silicon and germanium are two of the best-known semiconductors. To explain this, and many other similar cases, we need the band picture.

An electronic energy level in an isolated atom is very sharp, meaning that the electron in such a level always has just the same energy. If two atoms are brought into contact, this level is displaced, the amount of displacement depending on the interatomic distance. If the two atoms were identical, so that each had an identical energy level, the interaction of the atoms would result not only in a displacement of this level, but in a splitting into two levels, the separation between them increasing as the atoms approached. If many atoms are brought together to form a crystal, a similar effect occurs on a much greater scale. All the originally identical levels of the identical atoms will split up, so that there will be as many levels as there are identical atoms in the crystal; but the total splitting will be about the same that it was for two atoms. Thus, with the many atoms in the crystal, the group of levels will be practically continuously filled, and we call the result an energy band. Each of the original atomic energy levels will split into a band, and since these bands become broader as the atoms get closer together, it may well be that they will overlap into a continuum of levels; on the other hand, if the splitting is less, there may be gaps left between the bands.

We can now compare the number of levels in each band, and the number of electrons actually present in the crystal. We find that with valence and ionic compounds, as a rule, there are just enough electrons to fill certain lower lying bands, and leave higher bands empty. On the other hand, with a metal, certain bands are only partly filled, and there are empty levels immediately above them. It now proves to be the case that in an electric field, if the material conducts electricity, the field will accelerate the electrons, increasing their energy by very slight amounts. Thus, we can only get conductivity, as in a metal, when there are empty levels for the electrons to go into immediately above the occupied bands; in the valence crystals and ionic compounds, where there is a gap above the valence levels, we get no conductivity. All this is as in the atomic picture, but now an additional feature comes in which allows us to handle such cases as Si and Ge. In diamond, the gap between the valence level and the next occupied level is quite wide, and the material is a good insulator; but in Si and Ge, the gap is small. At high temperatures, electrons from the occupied valence band can be excited by thermal agitation to the upper, conduction, band, and then the material becomes a conductor; this is called intrinsic conductivity, and such conductivity increases rapidly with the temperature. But there is another mechanism for conductivity, which is more important. If there are impurity atoms in the crystal, their energy levels will not be incorporated into the main bands, for these bands come only from the interaction of the energy levels of identical atoms. Some of these impurity levels may lie within the gap between the valence and conduction bands. Examination then shows that either of two situations is possible. In some cases, a moderate temperature will excite an electron from one of these impurity levels to the upper, unfilled conduction band, thereby allowing conductivity; such an impurity atom is called a donor atom, and the resulting conductivity is called *n*-type, since the carriers are negative electrons knocked into the upper level. The other case is one in which an impurity level is normally unoccupied, and a moderate temperature will excite an electron from the lower valence band to this level, which is then called an acceptor level. The resulting conductivity is called *p*-type, since the carriers are holes in the valence band, which can be shown to have the characteristics of positive charges.

An interesting case intermediate between a conductor and a semiconductor is shown in graphite, an exceptional substance on account of its unusual crystal structure. In this case, the electrons just fill the lower valence band; but the next higher band begins immediately adjacent to it. The density of electronic states simply goes down rapidly to zero at the boundary between the bands, then immediately builds up again, being zero only at one point. This results in graphite showing properties intermediate between a conductor and a semiconductor, as far as its electrical behavior is concerned.

2. The Mechanical Properties of Solids

We have already mentioned that the mechanical properties of solids are well explained by the atomic picture, and that this handles such properties as compressibility quite satisfactorily. It is well known, however, that there are a number of mechanical properties which are structure-sensitive, such as strength, deformation resistance, hardness, and creep; and these properties cannot be understood without examining the effects of dislocations and impurities in the lattice. Perhaps the most striking of these effects is shearing strength. If we take the interatomic forces, and calculate the force necessary to make one atomic plane slide over another, and hence to shear the crystal, we find an extremely large value. On the other hand, if we take a single crystal and shear it, we find that the necessary force is perhaps only a thousandth of the calculated value. The explanation of this was given a number of years ago, and is something like the reason why it is so easy to tear a piece of paper, whereas the paper can support a large weight on a direct pull. Once a tear is started, it tends to propagate itself, since it is weakest just where it is tearing, and the whole force tearing it can be concentrated at that one point. Similarly in a shearing crystal, it is believed that so-called dislocations can start a particular type of atomic displacement which has this same effect of concentrating the shearing force at one point, so that its effect is greatly magnified, and a relatively small force can shear a crystal. The dislocation propagates itself across the crystal, as a tear develops across a piece of paper.

This theory of dislocations explains in a general way the ease of shearing a crystal, but it also leads to an understanding of hardening. Anything that can impede the motion of the dislocations across the crystal can stop the shearing action, just as a criss-cross of strong reinforcing strips across a piece of paper would strengthen it by stopping any tear when it reached a reinforcement. The simplest form of hardening is work-hardening, which results automatically as a crystal is distorted. It is supposed that, though each dislocation moves fairly freely through the crystal, still the slipping which it has produced leaves a somewhat damaged trail behind. As we have more and more of these damaged trails, we reach a point where a further dislocation meets one of these trails in its slip through the crystal, and is stopped from moving further. As more and more damage is produced, these frozen-in displacements impede the motion of any further dislocations more and more, and the material is work-hardened. By the time such dislocations are distributed through the lattice so close to each other that they are beginning to interfere seriously with each other, we approach a saturation situation, with a maximum amount of cold work, maximum change in physical dimensions, and a maximum amount of stored energy in the lattice, as we shall consider in a later paragraph. In a tensile curve of strain against time, a continual application of a constant force will then produce motion at a slower and slower rate, slow enough to that we can speak of it as creep.

Other mechanisms of hardening are hardening by alloying, and precipitation hardening. It is clear that anything which disturbs the continuity of the crystal lattice will impede the motion of dislocations and result in hardening; and an impurity atom, or an intrusion of of another material, will certainly have this effect. In solid solutions, the solute atoms are usually of appreciably different size from the solvent atoms; and when they are arranged at random, they disturb the regularity of the lattice, resulting in hardening. In precipitation hardening, the precipitate strains the matrix out of which it is precipitating, again disturbing the lattice regularity and leading to hardening.

These mechanical properties are closely tied up with temperature. A distorted lattice is generally unstable thermodynamically; if the atoms were free to migrate, they would take up positions of thermodynamic stability, and the lattice would become more regular, leading ordinarily to a reduction of hardness. This, of course, is the phenomenon of annealing. At a low temperature, the atoms are not free to move, because each one in general is held to a position of equilibrium by restoring forces; but it can well be that by displacing a relatively short distance in some direction, the force can reverse in sign, and pull it into a more stable position. The energy necessary to surmount the barrier is the activation energy, and to produce the irreversible transition to the more stable position the temperature must be raised so that the thermal energy becomes comparable with the activation energy; the higher the temperature, the larger the chance of the atom surmounting the barrier, and the faster the process goes. By measuring rates of annealing, one finds these activation energies in a well-known way, and finds often that different atoms in the distorted lattice have different activation energies.

Associated with the unstable positions of the atoms is the stored energy in the lattice. If an atom can go to a lower energy position, it loses energy in the process; it then has stored energy in its unstable position. There are two main ways of measuring this stored energy. One is to heat the material gradually; as we run through the temperature range where the atoms jump easily from their unstable to the stable positions, they will liberate an abnormal amount of heat, and we can measure this from the temperature rise observed. The other method is to measure the heat of solution; either a cold-worked or an annealed sample will take up the same final energy if it is dissolved, and since its original energy is different, the cold-worked sample having higher energy, the energy difference, measured in the heat of solution, must be different.

There are several other phenomena associated with atomic positions in the lattice, and their effects on mechanical and thermal properties, which are of considerable significance. Order-disorder transitions in certain alloys are rather well understood: below certain transition temperatures the atoms in such a substance as Cu_3Au, for instance, are arranged in an ordered structure, while above this transition the thermal agitation leads the atoms to be arranged at random, but on the lattice sites of the original crystal structure. This can be investigated by x-rays, by thermal measurement, and in some cases by electrical and magnetic methods. Another interesting effect is the damping of acoustical, or mechanical vibrations. Such sound waves show large absorption, or dissipation of energy, in certain ranges of frequency, which can be shown to be correlated to various irreversible processes in the crystal. Still another useful field of study is diffusion, either self-diffusion of atoms of one sort in a crystal of the same kind of atoms, or diffusion of impurity atoms. Study of activation energy of diffusion as it varies with temperature can throw considerable light on the processes by which atoms can change places in a lattice.

3. Electrical and Thermal Conductivity

The electrical conductivity of a material is made up of the product of two quantities: the number of carriers of current (as electrons in ordinary metallic conduction, or ions in ionic conduction), and the mobility of these carriers, or the velocity acquired by them per unit electric field strength. Let us consider how these quantities depend on temperature (the most conveniently adjustable physical variable) and on irregularities in the lattice, for the various forms of conductors.

In the first place, for a metal, the number of carriers is almost independent of both temperature and physical state of the sample, or is not structure-sensitive. It is of the order of one electron per atom. If the conduction band is not very full, the carriers act like negative electrons; but if it is nearly full, it is easier to think of the holes in the band as carrying current, and these act like positive charges. The number and sign of the carriers can be found from measurement of the Hall effect, the transverse voltage which appears in a conductor when a magnetic field is set up at right angles to a current. Unfortunately the Hall coefficient for metals is small, rather hard to measure, and its interpretation is not completely simple; its interpretation is in any case complicated when some carriers are positive, some negative. Thus, we do not possess as complete information about the number of free electons in metals as we wish we did.

The mobility of electrons in a pure metal is approximately inversely proportional to the temperature. The reason is simple, and informing. The electrons in the perfect crystal, as is shown by the band theory, act like waves, just as they do in the phenomenon of electron diffraction. If the lattice is perfectly regular, they travel through the crystal without scattering; the mobility would be infinite. They are scattered, however, by the irregularities of the lattice produced by temperature agitation. These irregularities are proportional to the temperature; and the resulting frictional force, inversely proportional to the mobility, is proportional to the temperature. In any pure metal other than a perfect crystal, there will be additional resistance, for there will be additional scattering of the electron waves by the lattice irregularities; and this will reduce the mobility, just as temperature agitation would do. It thus comes about that electrical resistivity is a sensitive test for lattice irregularity. This is particularly true of the resistivity at very low temperatures, for there the resistivity resulting from temperature agitation has disappeared, and any residual resistance is a result purely of lattice irregularities. Such residual resistance arises from cold work, and much greater resistance arises from the irregularity of crystal structure met in an alloy. It is interesting to note that in some substitutional alloys, in which the atoms are ordered in the lattice below a certain transition temperature, disordered above that temperature, the residual resistance is much higher above the transition temperature, in the disordered state, than below.

The resistance at very high frequencies, in the microwave region, has certain interesting properties which the dc resistance does not possess. At high frequencies, an electromagnetic wave penetrates a metal only for a short distance, called the skin depth; this depth decreases as the frequency gets higher, or as the conductivity gets higher, and the material becomes more opaque. The skin depth can well be as small as 10^{-4} or 10^{-5} cm. A measurement of conductivity by microwave methods then measures the value holding within this small distance of the surface. It is sometimes held that cold-working the surface of a metal leaves it with a quite disordered lattice at short distances from the surface; this microwave measurement then will be sensitive to this type of surface condition, a condition hard to investigate by most other methods. An additional interesting piece of information is yielded by this measurement, which does not come from low frequency work. The skin depth can become small enough, as the conductivity goes up at low temperature, so that it is smaller than the mean free path of the electrons, which increases as the temperature goes down and the conductivity goes up. As these two quantities cross, the whole mechanism of conductivity changes, and the measured conductivity does not continue to rise as the temperature is reduced. By observation of this effect we can get separate information about the mean free path in the surface layers, and hence about the mobility itself. Combined with the conductivity, this gives independent information about the number of carriers of current.

Next we consider the conductivity of semiconductors. We have already pointed out that the current carriers in such materials are produced at moderate temperatures from the donor or acceptor impurity levels, and at higher temperatures by direct excitation from the valence to the conduction band. This gives a characteristic dependence on temperature, which is the major feature resulting in the great temperature dependence of the conductivity of semiconductors; it can be easily investigated by the Hall effect, which is much larger than for metals. As for the mobility, it is rather like what we should expect by analogy with metals. There is a resistivity arising from temperature agitation, proportional to the temperature as with metals; and an additional resistivity, necessarily present, arising from the lattice irregularities coming from the impurity atoms. Very careful analyses of the conductivity and Hall effect of Ge and Si, carried out by Lark-Horovitz and the Purdue group, by the group at the Bell Telephone Laboratories, and elsewhere, have contributed a great deal to our recent increase of knowledge of the band structure and conductivity of solids.

For insulators, the situation is much like that for semiconductors, except for the absence of impurity atoms, and the greater width of the gap between valence and conduction bands. The materials would act like intrinsic semiconductors if we could heat them hot enough to excite electrons thermally to the conduction band; but generally the required temperature is so high that the material would melt first, or if it is an ionic crystal, ionic conduction would start. On the other hand, any mechanism which can excite electrons to the conduction band will result in conductivity, the mobility of electrons in this band being not particularly lower than in conductors. Examples are photo-conductivity, where the electrons are excited by absorption of light,

and bombardment-induced conductivity, as in diamond and other materials, where they are excited by incident high energy particles. As we shall see later, it is a characteristic feature of the passage of fast particles through matter that they excite many electrons to higher energy levels. In anything but an insulator, the number of electrons so excited is not great enough in comparison to those normally present in the conduction bands to make a significant change in the conductivity; but in a good insulator the number of conduction electrons normally present is practically zero, so that any excited electrons can be easily observed.

Thermal conductivity arises from two quite different mechanisms. In the first place, in a good conductor, the electrons can easily carry energy from point to point in the crystal. If they have an unusually large kinetic energy in one part of the crystal, corresponding to a high temperature, they rapidly carry this energy to the rest of the crystal. A well-known result of the electron theory of metals is that there is a constant ratio between this part of the thermal conductivity and the electrical conductivity, the so-called Wiedemann-Franz ratio, indicating that good electrical conductors are also good thermal conductors. The Wiedemann-Franz ratio, however, depends on temperature: at moderately high temperature, the thermal conductivity is roughly independent of temperature, while the electrical conductivity in a perfect crystal varies inversely proportionally to the temperature. Lattice irregularities will scatter the electrons just as much in their heat-carrying as in their current-carrying capacity, so that they will decrease this part of the thermal conductivity. In addition to this electronic thermal conduction, however, there is a quite different mechanism simultaneously present, and present to substantially the same extent in insulators as in conductors. This is conduction by the lattice vibrations. If an atom is set into vibration, it will act on its neighbors, set them into vibration too, just as would happen in a set of weights linked by springs. This type of thermal conductivity is a good deal smaller than the electronic conductivity and is not important in good conductors, but in insulators it is the whole effect. This lattice conductivity operates essentially by the spreading of sound waves, as the electron conductivity operates by electronic waves; and anything which scatters the sound waves will reduce this conductivity, just as anything scattering electron waves will reduce electronic thermal conductivity. We may expect, then, that it will be seriously reduced by lattice distortions.

There has been a fairly large amount of work on the theory of these electrical and thermal effects, but still it is far from enough. The cases of germanium and silicon are well understood, and some of the work on pure metals has reached a fairly satisfactory state. But the theory is far from the point where one can use observed measurements of conductivity as a function of temperature for an arbitrary material, and work back to say what must be the nature of lattice displacements to account for the observations. This is a result, not so much of basic inadequacy of the theory, as of lack of sufficient work to carry through the necessary calculations, and lack of experimental material to test the simpler predictions of the theory, so as to make its use in complicated cases more certain.

4. Types of Experimental Investigations of Solids

In the preceding sections, we have mentioned a number of types of experimental methods adapted to studying solids: mechanical experiments, such as stress-strain curves, measurements of hardness and creep; thermal measurements, as on liberation of stored energy on heating and on solution; electrical measurements, of the conductivity and Hall effect as functions of temperature; thermal conductivity as a function of temperature. We might add magnetic susceptibility, giving a certain amount of information about the electronic state of the atoms, provided they are magnetic, and giving an extraordinarily sensitive test for small amounts of ferromagnetic atoms in a ferromagnetic state. All these experiments give information about where the atoms are, in an indirect way; if we had theories well enough worked out relating these effects to lattice irregularities, we could use the measurements to determine the number and nature of such irregularities. Unfortunately, however, in almost every case the theory is not well worked out, and has not received enough experimental checks to allow us to believe its results without great reservations. What is needed is a great deal of further development of the theory of each effect, and many experimental measurements of several different effects on the same samples, so that if the nature of the lattice disturbances as deduced from the measurements of several effects agrees, we can begin to believe the theories of these effects. Such a correlation has been carried through, as we have mentioned earlier, for germanium and silicon, and the good agreement obtained there has given us confidence in the general correctness of the concepts of solid-state theory. It is time that similar work was done on a variety of materials of all different types.

In contrast to these effects, which give rather indirect evidence about lattice irregularities, there are several types of experiments which give quite direct information. Foremost among these is the method of x-ray diffraction, with its useful subsidiary tools of electron and neutron diffraction. The classical use of x-ray diffraction is to find the structure of a perfect crystal; but it can be adapted to studying crystal imperfections as well, and in the hands of such workers as Zachariasen at Chicago, Warren and Averbach at M.I.T., and Guinier in France, it is capable of giving much more detailed information about lattice imperfections than we have had up to the present. An x-ray diffraction study gives essentially a fourier resolution of the scattering power in a crystal; unfortunately it

gives only the amplitudes, not the phases, of the fourier components, so that we cannot synthesize these components and get the complete structure without making some hypothesis about the phases. It is in making these hypotheses that the good judgment and physical intuition of the crystal structure experts come in; they must work out structures consistent with known interatomic distances and such physical principles.

Now in a distorted lattice, in a similar way, we can get by x-ray diffraction a fourier resolution of the displacement of the atoms from their regular lattice positions, and here again we have only the amplitudes, not the phases. It is only very recently that such fourier resolutions have been obtained, and the synthesis of the components to give the lattice displacements is just starting, so that the workers in the field have not yet had enough experience to determine what are the reasonable guesses to make about phases, and the resultant types of possible displacements. There is good reason for hoping, however, that with experience it will be possible to synthesize the displacements, and get pictures of the irregularities in lattices almost as definite as our present pictures of the structure of regular crystals. This method, properly used, should prove to be the most valuable direct tool in investigating lattice irregularities. And as we have emphasized earlier, the positions of the atoms form in a sense the independent variable, the quantity in terms of which all other properties of the distorted lattice can be directly computed, whereas the opposite procedure, the attempt to deduce the positions of the atoms from such indirect evidence as electrical or thermal properties, is much less certain.

Just as x-ray crystal methods form the direct approach to the study of the atomic positions, and hence in a sense to the atomic picture of the solid, so there is another entirely different x-ray method adaptable to the band picture. This is the study of soft x-ray emission bands. If one of the inner electrons of an atom in a crystal is knocked out by bombardment with a fast electron, an outer electron of the same atom can fall into the hole so left, and the frequency of the emitted x-ray gives a measure of the energy difference between the two levels. If the electron fell into the hole from the valence band of the crystal, it could have come from any part of that band; and hence there will be a distribution of energies, and of frequencies of the emitted x-rays, giving direct information about the distribution of levels in the band (combined with the transition probabilities which can be estimated). This method, developed by O'Bryan and Skinner in the early 1930's, has been used to give band structures of a good many metals of low atomic weight, in general getting fairly good agreement with the theoretical expectation, and helping considerably to verify the theory. This method should be extended to additional metals, to verify it further and extend our knowledge of the band theory of these materials; but then it could in addition be used on alloys, compounds, and materials with lattice irregularities. We have already mentioned that it is the interaction of an atom with its identical neighbors which leads to the broadening of the bands. In a crystal where for one reason or another an atom is surrounded by different types of atoms, either because the material is a compound, or because the atom in question is an impurity atom or is in solid solution, the broadening of its band would probably be entirely different from its value in the pure metal. In a distorted lattice, the bands should broaden or narrow, depending on the type of strain. Many effects in these soft x-ray bands should then show interesting properties in alloys and cold-worked materials. These effects have never been investigated, but it might well be that the method could be developed into a valuable tool.

Closely related to these soft x-ray emissions is the method of photoelectric emission, as developed by Apker and others at the General Electric Company. Here, by absorption of optical radiation by the surface of a material, we raise electrons from the valence band to the conduction band. The resulting properties give information about the distribution of levels in the valence band, as well as about the width of the gap, which have shown good correlation with the results of the soft x-ray emission method. These two experiments, together with certain optical absorption experiments, and the study of the conductivity of semiconductors, as a function of temperature, are capable of giving detailed information about the band structure of solids. They have been applied so far to only very few materials, and the theory likewise has been worked out for only few materials. The progress of solid-state theory would be greatly aided by more work on these studies, both theoretical and experimental, of energy bands. For it is only when the nature of the bands is known that the theories of such things as electrical and thermal conductivity can be worked out with enough assurance so that they can be used for predicting the effect of lattice displacements on these properties.

II. THE NATURE OF RADIATION DAMAGE

In the preceding sections we have learned something of the nature of solid-state theory, with particular attention to the effects produced by displacements of the atoms. Some properties, like density, compressibility, and a few others, are relatively insensitive to displacements, whereas others, like electrical conductivity (particularly at low temperatures), thermal conductivity, and hardness, are particularly structure-sensitive. The effects of atomic displacements on these properties arise quite independently of how the displacements are produced. Thus, the properties and their theories can be studied by examining cold-worked material, and its annealing behavior, in which the dislocations are

produced by purely mechanical means. When such study has gone far enough, by the methods outlined in the preceding sections, we should be able to know quite definitely the type of displacement made by various forms of mechanical deformation, and the effect which this displacement has on structure-sensitive properties.

Radiation damage will produce displacements which in some ways are similar, in other ways undoubtedly different, from those resulting from cold work. It should be clear by now, however, that the methods of investigating it, both experimentally and theoretically, will be identical to those used for cold work. It thus becomes evident that any well-thought-out program of research on radiation damage should be combined with a program for studying cold work by the identical methods. By so doing, we shall have two independent ways of making atomic displacements, and we can not only study them separately, but can compare them, and in so doing throw more light on the nature of each phenomenon. A combined research program of this type is bound to advance not only our knowledge of radiation damage, but of the physics and metallurgy of the solid state as well, with its importance in the development of new materials of properties which are at present unattainable. Such fundamental work is at present sadly lacking, quite aside from the question of radiation damage.

When we consider the damage produced by radiation, we shall see that there are various types of incident radiation, which can act on the material in ways which in their broad outlines are similar, but which differ in detail. Just as a simultaneous study of cold work and radiation damage is more powerful than study of either by itself, so study of damage produced by different types of incident particles, as neutrons, protons, or deuterons accelerated in a cyclotron, and fission fragments, can furnish more information than damage produced by one type of particle only. If these different types of irradiation were considered only empirically, their comparison would not give much information, aside from an empirical conversion factor, telling how many times more effective one type of radiation was than another in producing damage. But in conjunction with effective theoretical guidance, to interpret the differences as well as the similarities of the various types of radiation, we could get more information than from any one type of radiation separately. A thorough study of this sort would allow us to extrapolate our knowledge to previously unused types of irradiation: for instance, to damage produced by a greater flux of fast neutrons than is attainable with present experimental facilities.

We shall now proceed to a description of the type of damage presumably produced by the various forms of incident radiation, the theory underlying the subject, and the types of direct experiment that might be used or have been used to verify the theory.

5. Theory of Damage Produced by Radiation

The important types of radiation for producing damage in metals and semiconductors, are neutrons on the one hand, and charged heavy particles on the other; the remaining sorts of radiation, photons or gamma-radiation and electrons, produce very little damage, except in insulators and chemical compounds, and mesons are not yet available in sufficient numbers for their damage to be of practical interest, though it would presumably be intermediate in effect between that produced by electrons and protons.

A neutron exerts almost no force on an atom unless it approaches close enough to the nucleus so that nuclear forces come into play; the remaining force, the magnetic force between the magnetic dipole moment of the neutron and magnetic fields present in the matter, is of some interest in problems of neutron diffraction and scattering, but is too small to be important in radiation damage. The cross section for neutron collision with an atom is then small, and the neutron range large: a neutron can traverse a considerable amount of material with only a few direct hits on atoms, at rather widely separated positions. These hits can then be handled as independent events. The theory of these collisions is simple, and well understood. There can be, in the first place, elastic collisions; these are the only important collisions with fast neutrons, and are the important ones for radiation damage in any case. In such a collision, the neutron gives a considerable fraction of its energy to the struck atom, the fraction being larger the more nearly the masses of neutron and struck atom are the same (that is, the lighter the struck atom). The rebounding neutron travels along to its next collision, where it gives up a similar fraction of its energy (therefore less in absolute amount), and so on, until it slows down. In the thermal range, large probabilities of producing nuclear reactions can of course come in, or if the struck atoms are fissionable, probabilities of fission.

The effect of the neutron on collision, then, is to produce either a recoil atom having large energy (often tens or hundreds of thousands of volts), or a recoil atom resulting from transmutation or fission. These recoil atoms in most cases will be charged, having left some of their electrons behind in the collision. We can then well start our study of radiation damage at this point, considering the damage done by the recoil atom, a charged heavy particle. This being the case, the study of neutron radiation damage is not different in principle from that produced directly by charged heavy particles: either by protons, deuterons, or alpha-particles from cyclotrons, or by fission fragments. We have, in other words, the problem of asking what happens when an incident heavy charged particle ploughs through a solid; if we investigate this question quite generally, as a function of mass, charge, and incident energy of the heavy particle, we shall have covered the whole theoretical background of radiation damage. The

effects of course, will depend both on the nature of the incident particle, and the type of solid under consideration.

Probably the most thorough study of this problem is that which has been made and is being continued by Bohr and his school in Copenhagen.* Their work has largely been concerned with the action of fast particles in gases, as in the cloud chamber, rather than in solids; but many, though not all, of the phenomena concerned are the same in a gas as in a solid, and concern only the interaction of the incident particles with the separate atoms struck, independent of the state of aggregation of those atoms. Let us then survey the behavior of fast particles traversing a gas. We first find that the fast particles themselves are likely to carry a charge, which is high for fast particles (like fission fragments at the beginning of their range), but decreases as the particles slow down. The fast particles acquire this charge, quite independent of whether they happen to have it when they start out or not. It seems fairly evident that if a very fast particle, as an atom, ploughs through a sheet of matter, its outer, more easily detachable electrons are likely to be scraped off in the process. A simple and fairly accurate rule of thumb tells which electrons will be removed, which will not. It is the following: all those will be removed whose velocity in their orbits is less than the velocity of the atom, while those with greater orbital velocity will remain attached to the atom. Since the outer electrons have small orbital velocities, the inner ones have great velocities; this means that the outer electrons will be removed, leaving a positive ion, but the number of removed electrons will gradually decrease as the particle slows down. The mechanism for this decrease of charge is a process by which the atom picks up electrons, rather than losing them, as it goes through the matter; it keeps in a sort of continual equilibrium, and those electrons which are just on the edge between being removed and captured are those whose orbital velocity is about the same as the velocity of the particle. Such equilibrium between ionization and capture has been known from the early days of nuclear physics, in its application to the loss and capture of electrons by alpha-particles passing through matter.

A fast ion passing through matter can have two quite different effects on the atoms it strikes. In the first place, it can make an elastic collision, giving kinetic energy to the recoil atom. This type of collision is not greatly different in principle from the elastic collisions of neutrons with atoms which we have already considered. Some of the results are the same: for instance, the energy given the struck atom is a larger fraction of the energy of the incident atom, the more nearly they have the same masses. But now, competing with this process is another one resulting from the charge on the ion: the incident particle can excite or ionize the electrons of the struck atom, resulting in an inelastic collision in which most of the energy lost by the incident particle is transferred to the removed electron, rather than to the recoil of the struck atom. The rate of this ionization can be shown to depend among other things on the square of the charge of the incident particle. Thus, it is clear that as the incident particle loses energy, and hence on the average loses charge, its rate of ionization will rapidly decrease.

When we take into account this decrease of ionization loss with decreasing energy, we find that at high energies all but a few percent of the dissipation of energy by the incident particle comes in ionization, while at low energies this ionization becomes unimportant, and the remaining energy is dissipated almost entirely in atomic recoils. Since we shall see that it is these atomic recoils which produce radiation damage, we see that a good deal of the damage of a very fast particle will in fact be produced at the end of its range, where it has been slowed down so as to move fairly slowly, though the integrated damage produced in the early part of the range can also be considerable. The energy where the situation changes over from ionization to recoil (for instance, the energy where half the dissipation results from each mechanism) depends on the type of incident particle, and is lower for a light particle than for a heavy one. The reason is that the condition is set really not by the energy, but by the velocity, of the incident particle; and the corresponding kinetic energy is then proportional to the mass of the particle. Thus, this transition energy for a proton is less than 10,000 ev, for an atom of intermediate weight (for instance a C atom) about 100,000 ev, while for a heavy atom, as a fission fragment, it can be over a million volts. For orientation, we note that these figures fit in with the corresponding situation for an electron, 1800 times lighter than a proton, in which inelastic collisions, with excitation or ionization, are produced at all energies greater than a few electron volts, whereas only elastic collisions, with nuclear recoil, are produced at lower energies.

We may now adopt the reasonable hypothesis that only that energy dissipated in nuclear recoil, which then can lead to atomic displacement, will lead to radiation damage; we shall come back to this hypothesis later. We can then draw some immediate conclusions about the damage produced by different types of particles. First we take a neutron. A fast neutron will produce recoil atoms of a few tens or hundreds of thousands of electron volts, at isolated points along its path. These recoil atoms will have intermediate mass (except in very light or very heavy elements), and hence will be in a range where they themselves can carry out further collisions, and where, moreover, they will dissipate perhaps half their energy in ionization, but the other half in recoils, producing secondary recoil atoms. These in turn will have considerably smaller energy, so low that

* The main outlines of the theoretical treatment are presented by N. Bohr in Kgl. Danske Videnskab. Selskab, Mat. fys. Medd. XVIII, 8 (1948).

they will hardly ionize at all, but will go on producing tertiary and further recoil atoms, of smaller and smaller energy. In other words, a considerable fraction, probably at least half, of the energy of each neutron recoil atom will go into atomic displacements. We shall consider later how much actual damage will result from this. The damage done by an individual neutron will then consist of a number of isolated regions of damage, each resulting from a single collision of the neutron with a knock-on atom; and a large fraction of the whole energy of the neutron will be potentially capable of causing radiation damage.

Next we consider a fission fragment. It is a heavy atom, starting out with an energy of the order of a hundred million electron volts. Roughly the first 97 percent of this energy is dissipated in ionization; only one or two percent of the energy is effective in producing damage until its energy is reduced to the order of a million volts. When it reaches this point, it will convert its remaining energy into recoil energy of knock-on atoms, which in turn will produce secondary and tertiary recoils as in the neutron case, so that the total potential damage done by a fission fragment is comparable to that done by a neutron of one or two million volts. The damage is very different in effect, however, because it is very much more concentrated. The collision cross section of a charged ion with an atom, in the region of elastic scattering, is very much greater than the cross section of a neutron for collision with the same atom. The reason is that the ion can act on the atom by electrostatic forces and ordinary interatomic repulsions, which have a very much greater range of action than the nuclear forces, which are the only ones effective between the neutron and atom. Thus, the fission fragment, instead of making elastic collisions with only a few isolated atoms, and producing isolated regions of damage over a range of perhaps a few centimeters, as a neutron does, will produce recoil atoms very close together, forming a practically continuous track of damage over a short range which, in a solid, can be as short as a few thousandths of an inch. The difference between these two forms of behavior is, of course, shown graphically in the cloud chamber and in the photographic plate technique for observing the tracks of fast particles; the first is a direct experimental means of investigating radiation damage in a gas, and the second in a solid.

An accelerated light atom, as a proton, deuteron, or alpha-particle, as produced in a cyclotron, will ordinarily start with an energy of ten or twenty million electron volts; and as we have seen, it will dissipate most of its energy in ionization until it reaches an energy of perhaps 10,000 ev. Accordingly, only the order of a few thousandths of its energy will be available for producing radiation damage. The recoil atoms from the primary collisions will have rather small energies, considerably smaller than the recoils from neutrons or fission fragments, and consequently there will be less effective production of damage by secondary and tertiary recoil atoms. The range of such a light accelerated particle, like the range of a fission fragment, is of the order of a few thousandths of an inch. An even larger fraction of the energy of such a light particle will be dissipated in electronic ionization, which by our hypothesis does not generate radiation damage, than in the case of fission fragments. We must remember in this connection that though this ionization energy is not effective in producing damage, it results in heating of the sample; and in any experiment on radiation damage this heat must be removed.

We shall now proceed to amplify these remarks in several different ways. First we shall give a critique of the present state of the theory, whose applications we have outlined, and of the ways in which it could be improved, together with types of experiment directly designed to test the theory. Next we shall give a discussion of the relative merits and comparisons of the various types of irradiation, for leading to a knowledge of radiation damage, taking account of the actual types of radiation sources available. Then we shall come to a more detailed study of radiation damage in solids, as distinguished from gases or isolated atoms, justifying our hypothesis that only the recoil energy, not the ionization energy, is effective in producing damage. Finally, we shall give some discussion of different types of solids, the type of damage to be expected, and a comparison with what is already known experimentally about them.

6. Critique of Present Status of Radiation Damage Theory

We have already indicated that a major contribution to the theory of radiation damage has been made by Bohr and his school. There are a number of other contributions to the subject in the unclassified literature, and classified reports by F. Seitz, H. M. James, F. W. Brown, and others. None of these, however, go far beyond the fundamental work of Bohr, though they carry Bohr's work much further in its detailed application to solids. The general situation of the theory is that in the range of very fast particles it is quite adequate, but when it deals with slow incident particles there is much further work to be done.

If two bare nuclei of a number of million volts energy approach each other, they will turn back only if they approach within a distance comparable to nuclear dimensions; the theory of such collisions depends on what is called the impact parameter, the distance of closest approach in a head-on collision, and under these circumstances this distance is of the order of 10^{-13} cm. It is clear that in this case the fact that the atom or ion is really surrounded by a cloud of electrons is almost immaterial. The time of collision, we find, is so short that the nuclei approach and recede from each other in a time short compared with the periods of the

electrons in their orbits, so that clearly they have no time to rearrange themselves or modify their shielding effects. Such a collision is then like the classical collisions of alpha-particles and nuclei in the Rutherford scattering experiment, and the theory of that experiment, which is one of the best-known in collision theory, is directly applicable. This is the part of the theory which is on really firm ground. The cross section for recoil in this range is a circle whose radius is of the order of magnitude of the impact parameter, explaining the very small elastic scattering cross section for fast particles.

In this same range of very fast particles, since the elastic scattering cross section is so small, it is clear that in most collisions the impinging nucleus will not hit close enough to the struck nucleus to produce elastic scattering, but instead will plough through the electron cloud. Here again the collision will be so fast that the electrons hardly have time to rearrange themselves during the collision. The effect of the collision in this case is to produce ionization. The method by which such a problem is handled is to analyze the very rapidly changing electric field of the incident particle, as it plunges into the atom and out again, into a fourier series in the time, as if it were a superposition of monochromatic light waves, and then to treat the effect of these monochromatic waves separately, as if we were handling a problem of photoelectric effect. This problem, though not so simple and straightforward as that of Rutherford scattering, still is a standard problem; and the solutions found for it are presumably quite accurate. In this range, we meet the first case where there might be a difference in behavior, depending on whether the target is gaseous or solid. The electrons knocked out of the atom, as we just mentioned, act much like photo-electrons, and are knocked up into ionized or excited energy levels. The arrangement of excited levels in a solid, where we have the band structure of levels that we have already described, is quite different in detail from that found in a gas, so that we may expect differences in probabilities of ionization between the two cases. Examination shows, however, that for the average energies given such an ionized or excited electron, of the order of ten or twenty volts, the differences between the gaseous and solid energy level distributions are not great, so that the results in the two cases are similar.

We may expect, then, that in this range of several million volt particles, the theory is in satisfactory shape. This expectation is verified by the fact that its agreement with experiments, performed both in Copenhagen and elsewhere, on the range and ionization per unit path length of fission fragments and other high energy particles in cloud chambers is good. More such experiments should be performed; there is an extensive program underway in Copenhagen, but this will hardly exhaust the possibilities of the field. And experiments should be performed on solids, for instance, on piles of thin foils, to see whether the corresponding predictions of range of fast particles in solids are borne out experimentally. But we are likely to expect here a verification of the theory rather than anything startlingly new.

The situation is quite different, however, when we come to slower particles; and unfortunately, as we have already seen, it is just in this region that the maximum radiation damage occurs, most of the energy being transferred to recoil atoms. The reason why the theory should be bad in this region is very clear. As the elastic scattering cross section becomes larger, and the ionization cross section smaller, this means that we are interested in collisions in which the nuclei do not approach nearly as close as they do in the high energy collisions. The nuclei may well not penetrate within each others' inner electronic shells. In such a case, we cannot handle the interatomic repulsions of the nuclei by themselves, as if the electronic shielding were not present. Bohr has made a first attempt to consider this shielding, but on the basis of a rather crude atomic model; Seitz has used a somewhat more refined method. We can see that in such a collision, taking a longer time than in a collision of very fast particles, the inner electrons may well make many revolutions in the time of a collision, and hence be able to modify their motion, leading to revised effects on shielding and on interatomic repulsions. With slow enough collisions, where the distance of closest approach is comparable with that met in solids under pressure, and the time of collision comparable with periods of rotation of the outer electrons, the electronic system may have time during collision to modify itself just as it would if a solid were compressed, or if two atoms in a molecule approached each other in the course of molecular vibrations. In such cases we know from the theory of interatomic forces what will be the nature of the repulsive interactions when the atoms approach each other, and we could consider the collision from this point of view; but it has not been done with any completeness.

The situation in regard to the theory, then, is that much additional work remains to be done in the region of slow particles; and this work should be checked by experiment. For these low energies, we may expect the situation to be quite different in the gas and the solid, since an atom in a solid, with a rather small energy, colliding with one of its neighbors, cannot cause that neighbor to recoil as if it were in a gas, because instead it will push it back into its neighbors, which will absorb the recoil energy like a cushion. We come here to characteristic problems of collisions in a solid, which we shall handle in a later section. It would be very much worth while, however, to do experimental work on such collisions in a gas, where it would be relatively easy, to check the theory, which is bound to be rather complicated. The experiments would not be in principle difficult: it would be necessary only to accelerate heavy particles to relatively low energies, a few hundred thousand volts or less, which could be achieved in

many types of accelerators, and investigate their elastic scattering in a gas, and their ionization of the gas. The results of such theory and experiment are just what are needed to fill in the most severe gaps in the theory, gaps which as a matter of fact render all the present estimates of number of displaced atoms in radiation damage very questionable.

7. Comparison of Various Types of Radiation, for Radiation Damage Experiments

In preceding sections, we have seen that the primary action of different types of radiation on solids is quite different, but that in the last analysis most of the damage is done by secondary and tertiary recoil atoms, after the primary particles have got slowed down to low enough velocities so that they are dissipating almost all of their energy in elastic collisions. This suggests that the type of radiation damage produced by different types of radiation will not be very different, and this supposition appears to be borne out by the available experiments. Some further work should be done to check this supposition; but if it proves to be correct, the question of which type of radiation to use in investigations of radiation damage becomes largely one of convenience.

The striking differences between different types of radiation damage are clear from the discussion we have already given. Neutron ranges are long, so that they will traverse samples of any ordinary size, and they will collide with isolated atoms throughout the interior of the sample, producing isolated centers of damage throughout the interior. Thus a sample irradiated in a reactor will show a volume effect, and large samples can be conveniently used. Particles accelerated in an accelerator will have small ranges, so that they will produce damage only very close to the surface of the material irradiated. The investigation of the effects produced by irradiation must then be done on a microscopic scale. Fission fragments will also have small ranges, and two cases can arise, depending on the disposition of the fissionable material. If it is dispersed throughout the sample, the fragments, though of short range, will be produced throughout the sample, so that the net result will be a volume effect; while if the sample is one not containing fissionable material, but placed in immediate juxtaposition to a sample emitting fission fragments, the results will be much as in cyclotron or other accelerator bombardment, concentrated close to the surface. A further difference between different types of bombardment, mostly from the practical point of view, comes from the amount of energy dissipated in ionization, and hence in heat, in proportion to the amount going into recoils. We have seen that in neutron bombardment a large fraction of the dissipated energy appears as energy of recoils, leading to radiation damage, whereas with both cyclotron and fission fragments only a small fraction appears as recoil energy, so that the heat removal problem can be very serious. This is particularly severe when the fissionable material is distributed throughout the interior of a sample, an arrangement which can be carried out in reactor irradiation, where the neutrons from the reactor produce the fissions; the practical problem of removing heat from the interior of a massive sample is formidable, since it must be done by heat conduction. The heat removal problem is inherently simpler in accelerator bombardment, where it is produced in essentially a two-dimensional region, so that the sample can be backed up with a massive material which is rapidly cooled. In all these cases, however, the temperature control problem is great compared with that met in direct neutron irradiation, where it is easy to keep the sample cool. This becomes a serious matter when we remember that radiation damage easily anneals out, the rate of annealing depending strongly on the temperature, so that an experiment carried out at elevated temperature, as we are likely to have with poor cooling situations, may give a very incorrect picture of what would be met with low temperature irradiation. Temperature control in radiation experiments, both with accelerator and reactor irradiations, is a very important technical matter.

Even though almost all the energy of an incident neutron is available to produce radiation damage, and only a small fraction of the energy of a particle from an accelerator, still the fluxes obtainable in an accelerator are so much greater than those available in experimental reactors that a little consideration shows that the radiation damage produced in that small volume of sample actually irradiated in the accelerator experiment is produced at a much greater rate than if the sample were in a reactor. Thus, there are practical advantages in using accelerator radiation. These advantages are partly balanced, however, by the fact that the reactor normally operates continuously for long periods of time, which most cyclotrons do not, so that long-time irradiations are simple; and that all experimental reactors have space for many more experiments than could be accomodated at one time in a cyclotron. For an experiment in which, for practical reasons of subsequent tests to be made on it, the sample must be large and uniformly irradiated, the reactor is the obvious tool to use. For fundamental experiments, on the other hand, there are advantages to both kinds of work. An obvious advantage of the accelerator is the possibility of using different types of accelerated particles, controlling their incident energies, and in general keeping the conditions of the experiment under control. In a reactor, where many experiments are going on simultaneously, and where the conditions are relatively inaccessible to the experimenter during operation of the reactor, careful control of conditions is much more difficult.

As regards the use of different types of particles of different energies, in accelerator experiments, more could be done than has been done in the past. Work so far has been done only with cyclotrons. We have

already seen that this is an inherently wasteful way to do the experiment, for a particle accelerated in a cyclotron, with ten or twenty million volts energy, dissipates only a very small fraction of its energy in radiation damage until it is slowed down to a few tens of kilovolts. All the rest of the energy goes into heating, is wasted as far as radiation damage is concerned, and is difficult to remove experimentally. There would seem to be much advantage, for this type of experiment, in using lower energy particles, say of a million volts or so, which could easily be produced in an inexpensive van de Graaff type generator. Such particles would produce very nearly as much damage per particle as would the particles from the cyclotron, and they would produce much less heat to be removed. Furthermore, in a van de Graaff generator it is easier to change from one type of particle to another than in a cyclotron. There would be much to be said for accelerating heavier ions, say argon ions or some such element, for radiation damage studies, since we have seen that for heavier ions the fraction of energy going into radiation damage is much greater than for a light ion, at energies in the neighborhood of a million volts. Of course, by starting with these low energy ions, the range of the particles would be correspondingly reduced; the high energy part of the range, in which the energy was being dissipated largely in ionization, would be eliminated. The experiments would have to be on an even more microscopic scale than in present cyclotron irradiations. But this should not be an insuperable obstacle. A single heavy ion, striking a target with a million volts of energy, should do nearly as much damage as a hundred million volt fission fragment, with very much less heating of the sample; in many ways, this might be a very desirable sort of particle to use for radiation damage study.

There is one aspect of radiation damage, which comes up conveniently in this connection, which we have so far omitted. The particle used as a projectile in every case will remain in the sample when it comes to the end of its range. If this particle were a knock-on atom from a primary neutron, it would be a type of atom already present in the substance, and would presumably appear as a displaced atom, but not an impurity atom. On the other hand, a fission fragment, or an atom shot in from an accelerator, will remain in the sample as an impurity unless it is a very thin foil, and after long irradiation there may be enough of these impurity atoms to change the properties of the material appreciably. This is a type of radiation damage which can be very important. If the impurity atoms are soluble in the matrix, the resulting solid solution will show increased hardness and electrical resistivity; with insoluble materials, it is clear that regions of impurity atoms will form, which can have many effects, on hardness, electrical resistivity, etc., of the type previously mentioned, just as important in some cases as the results of actual displacements of the atoms of the crystal. These impurity atoms can naturally be investigated, and their effects estimated, by using the same principles we have described above; but they must not be forgotten. In many ways the most striking effects may perhaps be expected from those inert gas atoms formed by fission, which will be completely insoluble in any solid, and might possibly form gaseous pockets within the solid, which could obviously affect its properties, particularly its mechanical strength, very adversely. The suggestion made above, of using accelerated argon atoms as projectiles for investigating radiation damage, would leave the argon atoms within the material, and might in this way simulate the effect of these fission particles in their effect on radiation damage.

8. Characteristic Features of Radiation Damage in Solids

So far the theory we have used in discussing radiation damage has not taken account of the actual nature of the solid state; we have used the simple rule that the energy of elastic recoil atoms went into radiation damage, while ionization energy of electrons did not. In the present section we shall examine this assumption more in detail, trying to get a definite picture of the mechanism of radiation damage, and hence of the probable state of the material in the damaged condition, and the consequent type of radiation damage to be expected.

We have seen that the characteristic problem to be investigated is that of a high energy charged particle ploughing through the solid; and for high enough energies, the limit depending on the type of particle, most of the energy will be dissipated in electronic ionization. Let us first see why we are justified in saying that this energy generally will not contribute to radiation damage. One way of justifying this is by the remark made at the beginning of our discussion, that the state of a solid is usually a unique function of the position of its atoms. This arose because the electrons can adjust their state so rapidly that they will almost immediately come to an equilibrium configuration. This suggests that the ionization energy of the few electrons excited by the incident particle will spread throughout the solid very rapidly, reducing itself to a low enough energy per unit volume so that it can be neglected. Another way to see it is to compare the two mechanisms of thermal conduction which we discussed in an earlier section. We remember that conduction by electrons was much more effective than by lattice oscillations. The ionized electrons will be able in this way to dissipate their energy, essentially by electronic conduction, very rapidly; they can do this even in an insulator, for even there the ionized electrons themselves are capable of carrying thermal energy. All this can go on in a time short compared with that needed for the atomic vibrations to dissipate energy.

Of course, this argument is in a sense specious. We know that general principles of thermal equilibrium

demand that energy originally given to certain degrees of freedom in a substance, as to the electronic motion, in time will become shared with all other degrees of freedom. Thus, this ionization energy eventually will be partly taken up by the atomic vibrations, the ordinary thermal energy. In fact, in a metal, for instance, almost all of it will eventually end up this way. For in a metal almost all the heat capacity resides in the atomic oscillations, almost none in the electrons; substantially the same thing holds for other types of solids. The question, however, is not one of eventual equilibrium, but of the kinetics by which this situation is reached. The method by which electronic energy can be transferred to nuclear vibrational energy is clear. Electrons in their motion can happen to be distributed, near a given atom, so that they are unbalanced, corresponding to more charge on one side of the atomic nucleus than on the other. This will result in an unbalanced electrostatic attraction on the nucleus, which will set the nucleus into oscillation, transferring energy from the electrons to the nuclei. The question is only how fast this process will occur, as compared with the process of spreading out of the electronic excitation through the lattice. This question has not been investigated theoretically as carefully as it should; but the indications are that the interchange between electrons and nuclei is slow compared with the spreading of electronic energy, so that by the time the nuclei take up the vibrational energy it is fairly well dissipated, so that individual atoms take up relatively small vibrational energy: an amount sufficient to result in the cooling problem we have mentioned in connection with cyclotron-irradiation, but not enough energy to produce radiation damage; for as we shall see, an atom must get a good many times thermal energy before it really is displaced from the lattice, producing permanent damage.

We have seen, then, our justification for thinking that the electronic energy generally does not lead to radiation damage. The only important exception to this generalization presumably comes in insulators. There, as we mentioned at the beginning of our discussion, it can be that the electronic conductivity is so poor that electrons can remain in non-equilibrium states for long periods of time. We can have several quite different effects presumably correlated with this fact. First we may mention the case of the crystal counter, for example the diamond counter. A high energy particle traversing diamond (for instance an alpha-particle of a million or so volts) will produce as its main result a great many electrons ionized to the conduction bands. If the diamond is in an electric field, these electrons carry current, which can be amplified, and allow the material to be used to count the incident particles. The electrons pile up in certain parts of the crystal, being caught in traps, and thus produce permanent damage of a type associated directly with electronic excitation, rather than nuclear displacement. Study of the number of electrons produced per incident particle would, as a matter of fact, form valuable verification of the general theory of ionization by fast heavy particles.

A similar example is found in some insulating crystals, as the alkali halides, where it is well known that electrons can be trapped in certain irregularities of the lattice, known as F centers, and can persist there for long periods of time, causing blackening of the material and other changes of physical properties. Such trapping of electrons can be brought about by exciting the electrons to the conduction band by passage of heavy particles, but also by other mechanisms: by the excitation of the electrons by absorption of light or gamma-rays, or by incident electrons. These are among the cases, mentioned in our introduction, in which electrons and photons can cause radiation damage. They can do it here, even though they can transmit their energy only to electrons with any appreciable efficiency, only because the materials are such good insulators that the trapped electrons can persist for long periods of time, without being neutralized by conduction, as would happen in an extremely short time in a good conductor like a metal.

A somewhat similar exception to the general rule that electronic excitation produces no radiation damage is found in the chemical effects of radiation. Many chemical compounds, particularly covalent compounds, have properties like poor conductors: an electron can become displaced from one covalent bond to another within the molecule, and it then has such a small chance of traveling back (as if the molecule were a poor electrical conductor) that effectively a permanent change is made in the molecule, covalent bonds are permanently modified or broken, and a chemical reaction occurs. This will happen provided the time taken for the electron to return to its original position is long compared with the time of the reaction. Most chemical reactions, as a matter of fact, which are affected by radiation have this sort of origin, in the electronic excitation rather than in nuclear recoil; though of course nuclear recoil also can have its effects in chemical processes. There is also good reason for thinking that the biological effects of radiation, such as the effects on mutation of genes, arise in the same way from a change from one electronic configuration to another produced by ionization, rather than from nuclear recoil. Here again the gene, effectively a very large molecule, plays the part of a very poor electrical conductor, so that a metastable electronic configuration may require an extremely long time to come to its equilibrium state; a mutation, in other words, may persist for a very long time or even permanently.

These cases of good insulators, in which electronic excitation can produce permanent radiation damage, are the exception rather than the rule in the sort of case usually considered. Most solids are fairly good conductors; and any electronic effect, as we have seen, can dissipate itself before it has any chance to produce permanent changes in the nuclear arrangement of the

material. We now come to the other aspect of the problem, the way in which nuclear recoils can lead to permanent displacements of the atoms from their regular crystalline arrangement.

We have seen that by the time a heavy particle moving through a solid has its energy reduced to a given level (ranging from the order of 10,000 volts for a proton to something over a million volts for a heavy atom) most of its remaining energy will be dissipated by elastic recoil. It will lose considerable fractions of its energy to a number of secondary particles. Each of those in turn becomes a heavy particle, of lower energy, moving through the lattice, so that each of these in turn will generate tertiary particles, and so on. We naturally ask, where will this process stop? And what will become of the particles which are hit, secondary, tertiary, and so on? There are two ways of considering this question, more or less complementary, leading to similar results. First, we may try to follow the history of each particle, primary, secondary, tertiary, and so on, and see how much energy each acquires, and how many collisions of a further generation it produces. We may estimate that an atom will be knocked out of its position in the lattice if it is given an energy greater than a certain minimum (estimated perhaps at the order of twenty-five volts), while if it is given less energy than this it will merely acquire elastic or thermal vibrational energy, without actually leaving its position of equilibrium and seeking a new location. In this way we can find the total number of displaced atoms. This is the method which has been used by Seitz, James, and Brown in most of the theoretical treatments of the subject.

A second approach to the problem is a more statistical or thermodynamic one. The nuclear kinetic energy is really heat energy. The process by which a fast atom knocks on its neighbors, transmitting its energy to them, is thermal conduction. What is happening is that we suddenly introduce a large amount of heat at a localized point of the lattice, and ask how this heat spreads out, and the rate at which the temperature falls. The kinetic energy not only of the primary atoms, but of the secondary and tertiary ones, and in fact of all those with energies of 25 volts and up, is so great as to correspond to an exceedingly high temperature, of the order of several hundred thousand degrees centigrade. It is clear, then, that a very considerable region of material around the track of the primary particle will be heated to a very high temperature. Thus if, for instance, the incident particle transfers an energy of 100,000 volts to recoils, then when this has become uniformly distributed among atoms each of 25 volts, we shall have 4000 such atoms; even when the average energy per particle is reduced to 1 volt, by interchange with its neighbors, so that we have 10^5 such particles, the energy still corresponds to more than the boiling point, even for a refractory material. There will be, then, a considerable track of high temperature, vaporized material about the track of a primary, often called a thermal spike. On the other hand, heat transfer is so fast, from such a small volume, or the rate of exchange of kinetic energy is so fast, that this region of excitation will very rapidly expand enough so that the material within it will have fallen below the boiling point, and will begin to approach room temperature. We have the equivalent of a very sharp thermal pulse, in a very small volume of material, followed by an extremely rapid quenching.

We can next ask what will be the effect of this rapid local vaporization and resolidification. If the process goes on in the interior of the material, as it will if the primary particles are recoils from neutron collisions in a reactor, the result will depend to some extent on the previous state of the region where the process occurs. If it was part of a perfect crystal lattice, it is highly unlikely that after melting and subsequent solidification it will remain as a perfect crystal. During melting it will certainly have become completely disarranged, and the freezing is so fast that there will be almost no chance for crystallization to form. Most likely, then, the result will be an aggregation of very small crystals, or even an amorphous material, bearing no relation to the crystal previously present. This will presumably persist, if the ambient temperature is low enough so that annealing takes place very slowly, as will be the case at room temperature for high melting metals, but not for low melting materials. If annealing proceeds rapidly at room temperature, the disturbed region will rapidly recrystallize, and essentially disappear. Thus, we must expect the effects of dislocations —hardening, increased electrical and thermal resistivity, and so on—in an irradiated material which has not been allowed to anneal, but essentially no effects if annealing takes place.

This type of argument suggests that there will be essentially a saturation effect with long enough irradiation. After all parts of the material have once had a chance to undergo the melting and recrystallizing process, additional radiation will not make further changes; each new disturbance formed will effectively take the place of one already there. A maximum hardness will be reached, comparable to the maximum attainable by work hardening, and a maximum change in electrical and thermal conductivity will be observed. This saturation state will be approximately the same whether we start with work-hardened or with annealed material, if we assume that the dislocations produced by radiation damage are not completely different from those concerned in work hardening. This type of argument would suggest that irradiation would not, for instance, increase the creep rate, since it would act like additional hardening. This conclusion might not be valid, however, as regards the creep under actual irradiated conditions. Suppose a material is under stress in a strong radiation field; at every instant a good many small regions will be locally melted, and the stress can

relax in those small regions. This will give a rate of relaxation of stress more rapid than would occur without the irradiation, and this is another way of looking at an enhanced creep rate. Whether this effect would be important, however, is not known.

We have spoken of radiation effects in an extended sample. At a surface, however, as at a surface of a cyclotron-bombarded sample, or under any circumstances at the surface of a thin sample such as a foil, other things can occur. If one of the regions of elevated temperature occurs at the surface, there is an obvious possibility that some of the material will boil off, and be lost to the sample. This is a process not unlike sputtering of material under cathode bombardment, in which the amount of sputtered material can be estimated by finding the amount which would be brought up to the boiling point by the action of an incident particle. This would suggest that a large part of the 10^5 or so atoms brought to the boiling point by an incident particle might boil away from the surface. This would probably be an overestimate, since some of these would certainly be so far inside the material that they would freeze before they could vaporize. Still, if we use such figures as this, we readily conclude that a foil, under radiation conditions, should disintegrate very rapidly. This does not look unreasonable from cyclotron experience, in which targets in the form of thin films deposited on backing material sometimes completely disappear in the course of a day or two, or from certain experiments on the disintegration of enriched U foils.

The effects of which we have spoken in this section are those of major importance in the study of radiation damage, and they are effects whose theory is particularly lacking. Much further work should be done, on the kinetics of the heating process, and of the cooling and recrystallization and annealing, of the local damage areas. Most of the theory which has been used so far has been of a very crude sort: trying to find the number of secondaries and tertiaries with energies greater than an assumed limit (say, 25 volts) required to displace the atom from its lattice. This is very crude in two respects. First, the collision theory is very inadequate for these slow particles, as we have seen earlier; secondly, there is no real justification for supposing that an arbitrary energy limit like this has any significance. It seems that the sort of picture based on a thermodynamic or statistical model, such as we have described, is closer to the truth, but this has not been worked out at all. Such a picture would have to be considerably more complicated than we have intimated. For instance, when a small volume of material was suddenly heated up in the interior of a solid, its pressure would go very high; and this would send out a shock wave, which would carry out some of the energy. The amplitudes of motion of the atoms in this shock wave would be so great that it could not be treated in a linear way; in fact, it is just the nonlinearity which would correspond to the possibility of permanent, irreversible displacements of the atoms. But this would be one approach. It will be a difficult task to work these things out completely, and they form perhaps the principal unsolved theoretical problem of radiation damage theory.

III. RADIATION DAMAGE TO DIFFERENT TYPES OF MATERIALS

From the principles which have been discussed in the preceding sections, we can make some guesses as to the type of damage to be expected with different types of materials. In most cases some preliminary experiments have been carried out, which often throw light on these expectations. We shall now survey the various types of materials, giving a very brief summary of expectations and findings for each, as well as suggestions regarding further useful types of work to be carried out. We shall arrange our discussion in the order of decreasing metallic properties: metals (including pure metals and alloys); semiconductors; insulators; covalent molecular materials.

9. Metals

From the survey we have already made, we should expect the effect of radiation damage in metals to be similar to cold work. That is, we should expect it to increase the hardness, since it would introduce dislocations inhibiting shearing deformation. These dislocations would scatter electron and thermal waves, so that irradiation would increase electrical and thermal resistivity. The total possible change should, however, be comparable to the total effect produced by cold work. Thus there should be a saturation effect. Furthermore, radiation should produce less effect on a cold-worked than on an annealed sample, since the cold-worked sample already has many dislocations, which would merely be replaced by others under irradiation. Saturation of either cold work or radiation should occur when substantially the whole of the material has undergone as much deformation as it can sustain. Furthermore, the amount of these effects shown by a particular metal should depend strongly on whether it is a metal which anneals rapidly at room temperature, in which case essentially no results of radiation damage should be observable, or if it is a metal which does not anneal, in which case radiation damage effects comparable to those of cold work will be found, but will anneal out at elevated temperatures.

These expectations are borne out by the observed experimental results. Materials such as Al, which anneal at room temperature, show practically no observable effects of radiation damage, whereas Cu, for instance, which does not, shows hardening under irradiation, comparable with the maximum change shown by cold work. It would contribute to our knowledge of these effects if more fundamental work were done on the determination of the type of atomic dislocations produced both by cold work and by irradiation (using for

instance x-ray methods), correlated with the related changes of hardness and of resistivity, to establish the correctness of the hypothesis that the same type of dislocations will always lead to the same physical results, no matter how produced, and to get quantitative measures of the maximum amount of dislocation which can be built up in a metal. This should be coupled with annealing studies, to determine heats of activation for the annealing process. Those metals which anneal at room temperature could be investigated by carrying out the experiment at reduced temperatures.

The situation with alloys is different from that with pure metals, in that phenomena such as disordering of an ordered structure and precipitation hardening, can be found, and can be affected by radiation. A limited amount of work has been done on problems of this sort, enough to show the type of result to be expected, though far from enough to understand the situation completely. Thus, in the matter of disordering, it is well known that the alloy Cu_3Au is ordered below a certain temperature, disordered above. When the ordered alloy, below the transition temperature, is irradiated, it becomes disordered, as is shown by increases both in the hardness and in electrical resistivity. The maximum hardness and resistivity are similar to those of the disordered alloy; and these changes persist indefinitely, so long as the alloy is kept below a temperature at which annealing occurs. It would be highly desirable, though this has not been done, to make x-ray measurements of the degree of order with different amounts of irradiation, and make direct correlation of this degree of order with resistivity, hardness, and other physical properties.

An example of a precipitation-hardening alloy is the Be-Cu system. Here more Be can be dissolved in Cu at high temperatures than at low, but by quenching from the high temperature the solid solution can persist to room temperature with a composition characteristic of high temperature. Heat treatment at a somewhat elevated temperature causes the excess Be to precipitate out in the form of BeCu, and the resultant intrusions of the precipitate in the solid solution cause hardening, and increase of resistivity. Irradiation of the alloy, under various degrees of heat treatment, shows in every case further increases in hardness and resistivity, which approach a saturation limit roughly equivalent to the maximum values reached by precipitation hardening. It is not entirely clear from the results so far whether this effect of radiation comes from dislocations of the lattice, or from precipitation induced by the radiation damage, but it seems plausible that the latter effect should occur, at least to some extent, since local regions of the crystal would be heated, and heating under the supersaturated conditions existing would allow precipitation. On the other hand, the cooling from the state of radiation damage corresponds to a very rapid quenching, as we have already indicated, so that perhaps the material would quench without further precipitation. Further work on this sort of case, including the very important similar case of stainless steel, is certainly needed, and might well be combined with theoretical investigations, to try to elucidate the thermal state of the sample during irradiation.

None of the effects of radiation on metals which we have mentioned are very striking, or likely to affect the usefulness of metals as structural materials in the reactor; the relatively small changes in properties to be expected can be allowed for in the design. There are several possible effects of radiation on metals, however, more difficult to consider in a theoretical way, but potentially more important practically. These are the possible enhancement by radiation of the creep rate, the rate of diffusion, and the rate of corrosion. These would all be of great significance in materials used as structural elements, or to contain fuel and fission products, or coolants, in connection with a reactor.

From our previous discussion, it is clear that we should not expect irradiation to increase the creep rate of a metal after irradiation; but the possibility exists that creep might be more rapid during a heavy irradiation, since the irradiation might keep enough of the atoms of the material stirred up so as to allow stress to relax more rapidly than normally. This same stirred-up condition would naturally favor any process demanding that the atoms move around, and diffusion and corrosion are two such processes. It is not known, either experimentally or theoretically, whether these effects would be great enough to be serious. This is the type of question where theory is not likely to be reliable enough to be trusted, so that experiments on these questions become very important from a programmatic point of view. Fortunately, a number of experiments bearing on these matters are being planned.

A very special case of radiation damage in metals is that arising in samples containing fissionable material. Here the damage may be expected to be greater than in a similar solid containing no fissionable material, but subjected only to neutron bombardment, for several reasons. First, a neutron inducing fission liberates a large amount of nuclear energy within the sample, so that the total store of energy potentially capable of producing damage is much greater than that supplied by the neutrons alone. Secondly, while the range of a neutron is fairly long, so that ordinarily it produces only a part of its damage in any small volume, the range of a fission fragment is small, so that all the damage done by any fission is concentrated in the immediate neighborhood of the fissioning atom. Thirdly, the fission fragments will remain within the sample, acting as impurity atoms, and consequently affecting the physical properties of the material, particularly if they are gaseous or other insoluble substances.

The general expectations which we have mentioned regarding radiation damage in substances containing fissionable materials, generally intended as fuel elements for reactors, seem to be borne out by observation;

but the experimental information is meager, largely on account of the extreme experimental difficulty of making observations on fissionable material after irradiation, a consequence of its strong radioactivity. Probably the work on radiation damage to fissionable materials will have to be largely programmatic for the present, devoted to answering the question as to whether particular forms of fuel elements have large damage or not, on account of the experimental difficulty of the problem. As more programmatic work is carried out on different types of fuel elements, however, using shielded laboratories for experimental investigation, it will undoubtedly be possible to gain increased fundamental understanding of the behavior of uranium and its alloys under irradiation.

In connection with all these problems of damage in metals, but particularly in fissionable materials, we must recall what was said in an earlier section, about the contrast between damage in bulk materials and in materials with a large surface-volume ratio, as thin foils. We recall that in the latter we may expect a very considerable vaporization from the surface, leading to disintegration of the foil in a rather short time.

10. Semiconductors

The principal elements which are semiconductors are Si, Ge, Se, and Te; in addition, a good many compounds, such as Cu_2O, have similar properties. Of these, Si and Ge have been studied very extensively, on account of their importance in the electronic field as rectifiers and more recently transistors. We have already mentioned the excellent checks of theory which the groups at Purdue and the Bell Telephone Laboratories have found for the way in which the number of current carriers, and the mobility of these carriers, depend on the temperature and the impurity content. They form, then, particularly good materials to use for testing our fundamental understanding of the nature of radiation damage; and it is rather disconcerting, but an encouragement for much further work, that at least two entirely unexpected and, so far, unexplained phenomena have appeared in the work so far carried out.

Germanium can be prepared with either donor or acceptor type impurities, so as to be *n*-type or *p*-type in its properties. Now it is found, when either type of Ge is bombarded, that the effect seems to be to create additional *p*-type centers. That is, when *n*-type Ge is bombarded, the number of negative carriers at first decreases, so that the resistance goes up, as if the *n*-type centers were being neutralized by new *p*-type centers; but as the irradiation is increased, the number of *p*-type centers continues to increase, so that the sample becomes *p*-type, its resistance again goes down, and it behaves in all respects like a *p*-type sample. On the other hand, when *p*-type Ge is bombarded it simply seems to get additional *p*-type centers. This behavior is comprehensible if we assume that the hole left behind when a Ge atom is knocked out of its place in the lattice acts like a *p*-type center; and there is some independent reason for thinking that this is the case. This behavior of *n*-type germanium on bombardment leads to the possibility of making a sample with a *p*-*n* boundary by irradiating part of a sample, thus obtaining rectifying properties without the need of any point contact.

All this seems to be in good order; but as we have mentioned earlier, two very serious paradoxes appear when we examine the situation more carefully. The first comes from the actual amount of damage produced. The work done previously, on unirradiated materials, has shown very accurately the number of *p*- or *n*-type impurities present in any sample, and the effect they have on conductivity. We can then deduce the number of *p*-type centers produced by irradiation, by measuring the change in resistivity. The remarkable fact emerges that, rather than having many centers for each neutron or deuteron traversing the sample (the experiments have been made both with cyclotron and reactor irradiation), as we should expect from the general picture of radiation damage presented earlier, we seem to have on the average only about one center per particle traversing the sample. This remarkable fact needs much further work to elucidate it; it may, however, indicate that annealing or healing of the damage in Ge is much faster than we should expect, so that most of it has disappeared before the observation is made. This does not fit in well with other observations, however. Under electron bombardment, Ge shows some damage, which anneals out at a very low temperature. The damage produced by neutrons or deuterons, however, only anneals out at a fairly elevated temperature. It would be very remarkable if most of the effect of the incident particles annealed out at room temperature, so that it was never observed, but that the residual effect needed such a high temperature to anneal it.

The other paradoxical situation about Ge is the comparison of its behavior with that of Si, and of other semiconductors. In all previously observed properties, Ge and Si have shown very close similarity. When irradiated, however, Si behaves completely differently from Ge. When either *n*-type or *p*-type Si is irradiated, its resistance in all cases increases. Furthermore, this is found with all other semiconductors. This in itself is a rather surprising result. The effect is very large, so large that good insulators can be made from the materials; and this in itself may prove to have practical importance. But it is such a large effect that it cannot be ascribed to decreased mobility of the carriers caused by lattice irregularities, but must instead arise from a decrease in the number of carriers themselves. Somehow the radiation must produce traps which capture the electrons or holes, destroying the conductivity. This in itself is surprising enough, but the particular difficulty is that of seeing why Ge and Si should behave so differently.

These curious features of the behavior of Ge and Si show that much further work must be done to understand the behavior even of these two materials, which in other respects are better understood than any other solids, under irradiation. Active work is underway at Oak Ridge, Purdue, and the Bell Laboratories on this problem, but even here additional groups might well enter the field. For instance, x-ray studies, and soft x-ray emission bands, would throw further light on the band structure and lattice irregularities. These materials are so uniquely suited to careful study of a variety of properties that further work on them, both from the standpoint of radiation damage and of their other properties, is very desirable.

11. Ionic Compounds and Ceramics

In this section we may group together a group of insulators including the alkali halides, typical ionic crystals, and metallic oxides, carbides, nitrides, and such compounds, which have some of the properties of ionic crystals, some of valence compounds. Some elements, of which diamond is a good case, likewise are insulators. These materials have large energy gaps between their valence and conduction bands, so that ordinarily they have no conduction electrons, and an exceedingly low electrical conductivity. Their thermal conductivity is only by lattice vibrations. They tend to be brittle at fairly low temperatures, though as they are heated up they become more plastic. The effect of radiation on the alkali halides has been an important subject of research for many years. The work has been mostly with electronic bombardment or gamma-radiation; as we have mentioned earlier, radiation damage by these types of radiation can be expected in insulators, though not in conductors. The effects are very involved, and we can only mention their nature. On irradiation, certain so-called color centers are formed in the material, causing it to become colored or opaque. These are believed to arise from electrons which become trapped in lattice vacancies occurring normally in the crystal. There are a variety of such color centers, with different activation energies for annealing, and different types of optical absorption; and on irradiation of these centers with light of various wavelengths, they undergo further changes, other centers appearing. This very involved situation, though it has been studied as we have said for many years, still is not completely understood. Work by Pringsheim at the Argonne has shown that neutron irradiation produces damage in the alkali halides of essentially the same sort, though there are some differences as compared with gamma- or electron-irradiation. Further study of these phenomena should be continued, for its importance in the fundamental knowledge of the theory of ionic crystals.

The behavior of the alkali halides under irradiation may well give hints of what to expect with oxides, carbides, etc., such as BeO, MgO, Be_2C, and similar compounds of the other alkaline earth and other elements, in which no such complete study has been made. These materials are heat-resistant ceramics suitable for use at high temperatures, and likely to be used in high temperature reactors. These materials have not been studied nearly as much from a fundamental point of view as the alkali halides, and only BeO among them has been much studied under irradiation. The thermal conductivity decreases, as is natural from the scattering of thermal waves by the lattice irregularities; this decrease is large enough to result in large thermal stresses under conditions of large temperature gradient, seriously interfering with the use of this material. Also, decreases in Young's modulus and breaking strength are observed, which may be a result of the presence of interstitial atoms, leading to a weakening of the lattice. The damage in ceramics appears to be considerably larger than in most metals. This is possibly associated with the ability of metals to flow plastically and thus relieve excessive internal stresses, whereas ceramics, being brittle, cannot relieve their stresses in this way under ordinary conditions. Some reservations must be made, however, about conclusions regarding the extent of radiation damage to such materials when irradiated at very high temperatures, drawn from experiments made on relatively cool samples. It is always possible that in these materials there is more tendency toward plastic flow at high temperatures, so that the mechanical properties may be better under high temperature irradiation than at lower temperatures.

Many different types of experiments might well be done with these ceramic materials, to increase our fundamental knowledge of their behavior. There should be x-ray measurements, stored energy, electrical resistivity, and Hall effect as a function of temperature over a wide range before and after annealing subsequent to irradiation. Interesting experiments on mechanical strength could be made not only at high temperatures, as we have suggested, but at high pressures. At high pressures, many ceramics (e.g., TiO_2) become appreciably more plastic, sometimes elongating as much as 40 percent before rupture in a tensile test. Accordingly, it would be of some interest to study radiation damage in such materials with and without the presence of large hydrostatic pressures. The pressures required are of the order of 20,000 or 30,000 atmos. Optical experiments, such as those performed on the alkali halides, should be undertaken also with the oxides, carbides, etc. There is some preliminary indication that interesting results would be obtained. In connection with optical studies, it seems desirable to supplement this work with studies of the type made by by Breckenridge on the power factor and dielectric constant as a function of temperature and frequency. Such experiments give additional information on the number of vacancies generated. On the other hand, irradiation of ionic crystals appears to afford a way of providing samples with a sufficient number of vacancies so that research

on diffusion rates in ionic crystals may be rapidly accelerated. A further type of experiment which should give useful information about the band picture, both of alkali halides and of ceramic materials, would be the soft x-ray emission study of their electronic bands. Still another type of experiment likely to produce interesting data would be the study of electrical breakdown with and without radiation. Such experiments should discriminate between the various current theories of this phenomenon.†

Taken altogether, we see that there is a great deal of work to be done in the field of insulators, both from the standpoint of solid-state theory and concerning radiation damage, and that the results to be expected here may well have more variety and interest than with metals.

12. Molecular Solids

We have already seen that radiation damage in a molecular compound is likely to be of quite a different nature from that found in most other solids: it is likely to be a result of the electronic ionization or excitation, rather than the nuclear recoils, produced by the radiation, and hence is likely to be similar whether produced by electron bombardment, gamma-radiation, or heavy ions. The field of radiation damage to chemical substances is, of course, one of great richness and variety, on account of the enormous number of types of chemical substances. As far as practical applications to reactor design are concerned, the most important materials are plastics of all types, and lubricating oils; in general, these materials show very serious and striking radiation damage, varying greatly from one material to another, for reasons which so far are not very well understood. Leaving for a moment the field of solids, there is also an important chemical problem in connection with the use of water in reactors: water itself undergoes dissociation under the action of radiation, which renders it much more highly reactive than in its ordinary state, so that chemical reactions and possibly corrosion can be accelerated by the action of radiation.

The primary mechanism of radiation damage in molecular substances is, as we have mentioned, the formation of ions or excited atoms in the material. The probability of their formation is a part of the fundamental problem of studying the ionizing effect of impinging particles on matter. Once they are formed, the rest of the process is one of reaction kinetics, involving successive transformations of the ions or excited atoms, formation of free radicals, eventual combination of the free radicals to form new chemical substances. Study of these processes will throw light on other problems of reaction kinetics, and, on the other hand, will be aided by any improvements in the knowledge of kinetics. Such studies in gas reactions are often carried out by use of mass spectrometers, to observe the various products of a reaction, and study of gas reactions forms a very useful way of approaching the general problem of radiation damage in molecular substances. In liquids and solids, the phenomena are more complicated, on account of the very short time that a free radical can ordinarily survive before it becomes stabilized by collision with another atom or molecule. There are few generalizations that can be made on the basis of experimental work to date on radiation damage in molecular compounds. One generalization regarding organic compounds which appars to be true is that aromatic compounds seem to show less damage than aliphatic compounds, the benzene ring somehow being able to absorb considerable electronic energy without dissociation.

There is one aspect of radiation damage to chemical compounds which is so straightforward that one can recommend additional work. This is the action of radiation on a photographic plate, such as is used in detecting cosmic-ray particles. It is well known that the tracks of particles, the number of ions produced per unit length along the path, and such things can be studied microscopically. Ordinarily the plate is used simply as a tool to investigate problems in the physics of high energy particles. It would well repay study, however, to reverse the process, regard the plate as a unique solid in which radiation damage can be made visually observable, and make careful experiments on the damage made by particles of different mass, charge, and energy, studying as far as possible the behavior of nuclear recoils as well as ionization.

† For a number of the suggested experiments in this paragraph, the writer is indebted to A. W. Lawson, Jr.

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

Contributed Original Research

Electron Optical Exploration of Space Charge in a Cut-Off Magnetron

D. L. REVERDIN*
National Bureau of Standards, Washington, D. C.†
(Received June 12, 1950)

The main effort in the work described has been directed towards the measurement of the electric field distribution in a cut-off magnetron, based on the electron optical shadow method. The space charge distribution can be determined qualitatively from the measured electric field distribution. The first objective is the development of a convenient method for such a study. In this work an elementary magnetron was investigated by sending an electron beam axially through the tube.

The conclusions of the study are that the space charge distribution does not conform to Hull's classical theory (which predicts that the space charge density is approximately constant out to a well-defined radius and zero beyond this cloud radius). Three different space charge distributions have been observed. One distribution has a sharp maximum near the cloud radius. Another falls off gradually toward the cathode; a third is time-variant. The shape of the space charge distribution is closely related to the symmetry of the magnetron.

I. INTRODUCTION

THE magnetron problem has been investigated theoretically by numerous workers. Unfortunately, because of the formidable mathematics involved, no exact solution can be obtained. The various simplifications that have been employed lead to inconsistent results. The classical theory of Hull,[1] further developed by Brillouin[2] and Megaw,[3] predicts that the space charge density is approximately constant out to a well-defined "cloud radius" and falls abruptly to zero beyond this radius. Therefore, the classical magnetron theory cannot explain the experimental fact of the presence of a small anode current above cutoff, called the excess anode current. An experimental investigation of the true physical conditions inside the magnetron is of fundamental importance for the understanding of the operation of the magnetron and might lead to eventual improvement of the present tubes.

Experimentally, little work has been done on the investigation of the space charge and the orbits of the electrons in a magnetron. To date, the only direct measurements of space charge fields inside a magnetron are those reported by Engbert.[4] His method consists of inserting a thin tungsten wire probe into the magnetron. Although it is possible by this method to measure the potential distribution in the tube, the very critical symmetry of the system may be disturbed.

A method for the mapping of otherwise inaccessible electromagnetic fields, called the electron optical shadow method,[5] has recently been developed. This method, when slightly modified for the exploration of the electric field inside a magnetron, circumvents many of the difficulties encountered in Engbert's method.

II. ELECTRON OPTICAL OBSERVATION

2.1. Description of the Method

The shadow method makes use of an electron lens system to produce a shadow image of a fine wire mesh placed in the path of the electron beam. The principle of the method and the mathematical details can be found in the literature.[5]

The modification used in this investigation makes use of two wire screens as shown on the diagram of the method in Fig. 1. The elementary magnetron (consisting of an axial cathode, a concentric cylindrical anode, and an axial magnetic field) is placed axially into the viewing beam. The field and space charge distribution inside the magnetron are inferred by the deflection effects they produce on the electron beam going through the magnetron. To measure these deflection effects we observe the shadow patterns of two wire screens W_1 and W_2. W_1 (see Fig. 1) is placed just before the magnetron. The function of this wire screen is to define the radial distance of the electron trajectories entering the object

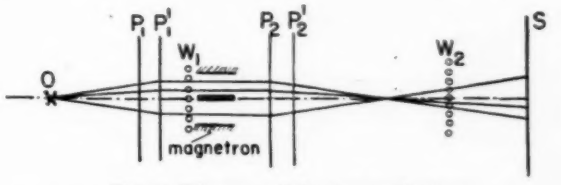


FIG. 1. Diagram of shadow method.

* Formerly Guest Worker, Electron Physics Section, National Bureau of Standards; now at the Centre de Recherches, Compagnie Générale de TSF, Paris.

† This paper is part of the doctoral dissertation submitted by D. L. Reverdin to the George Washington University, Washington, D. C., in February, 1950.

[1] A. W. Hull, Phys. Rev. **18**, 31 (1921). A. W. Hull, Phys. Rev. **23**, 112 (1924).

[2] L. Brillouin, Phys. Rev. **60**, 385 (1941).

[3] E. C. S. Megaw, J. Inst. Elec. Engr. (London) **72**, 326 (1933).

[4] W. Engbert, Hochfrequenztechnik und Elektroakustik **51**, 44 (1938).

[5] L. Marton and S. H. Lachenbruch, J. Appl. Phys. **20**, 1171 (1949), where other references are listed.

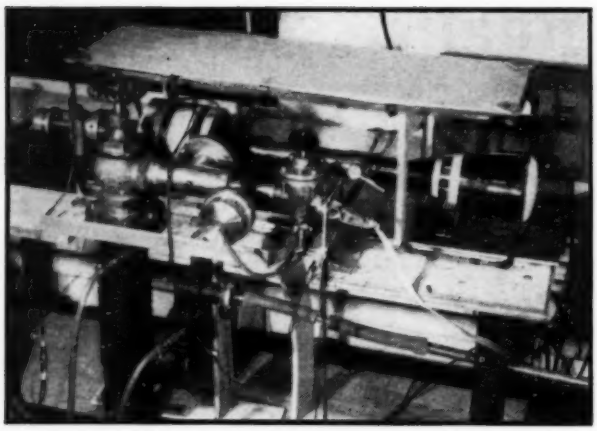


FIG. 2. Arrangement of apparatus.

(i.e., the magnetron) space. A second wire screen W_2 is placed behind the axial crossover of the trajectories. These two screens serve as coordinate systems to measure the deflections suffered by the various portions of the beam in going through the magnetron. The magnetic lens effect of the Helmholtz coil, which produces the magnetic field, is indicated on Fig. 1 by two pairs of principal planes P_1P_1' and P_2P_2'.[6] The lens effect is produced by the field on both sides of the constant field region, which acts like two identical lenses. The strength and position of these lenses is such as to produce electron paths approximately parallel to the optical axis in the object region. The object (the magnetron) appears on the fluorescent screen S as a shadow.

When this system is illuminated with the viewing electron beam from a point source O, a superposition of shadow patterns from both wire screens is seen. Only the inside contours of the anode and the cathode supports of the magnetron are visible.

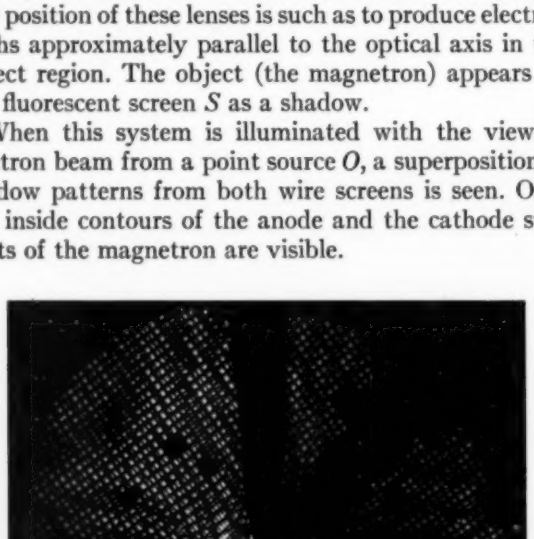


FIG. 3. Pattern obtained with electric field.

[6] These principal planes are shown as flat planes for simplicity. In reality, they are curved; but the investigation method is independent of this curvature as well as of all other aberration effects.

The shadow pattern on the fluorescent screen appears differently in three cases: (a) no electric field in the magnetron, (b) electric field without space charge, (c) electric field with space charge. After the apparatus has been calibrated by the measurement of the electric field distribution without the space charge (from the patterns (a) and (b)), the technique consists of taking pictures of the fluorescent screen and studying the difference between patterns (a) and (c) from which the space charge distribution can be computed.

The experimental arrangement had to be flexible and is designed as an electron optical bench shown in Fig. 2. The use of a stroboscopic technique[7] made it possible to observe the magnetron while no heating current flowed through the cathode.

2.2. Computation of the Space Charge

For a magnetron without space charge the electric field E, directed inward along the radius, varies as $1/r$, and is expressed by

$$E=K/r, \quad \text{where} \quad K=V_a/\log(r_a/r_c). \tag{1}$$

V_a is the voltage of the anode; r_a is the anode radius; r_c is the cathode radius. The presence of the space charge will change the electric field distribution so that the field varies more gradually than indicated by formula (2).

The unknown electric field at a particular radius can be obtained from the equation relating the field strength E of a uniform electric field to the small deflection angle θ' produced by such a field. This is

$$E=2\theta' V_0/z_0, \tag{2}$$

where V_0 is the voltage with which the electrons of the viewing beam are accelerated in the electron gun, and z_0 is the length of the electric field in the direction of the optical axis. This analysis makes two assumptions: (1) that the angle of deflection is so small that the radial

FIG. 4. Pattern obtained without electric field.

[7] L. Marton and D. L. Reverdin, J. Appl. Phys. **21**, 617–618 (1950).

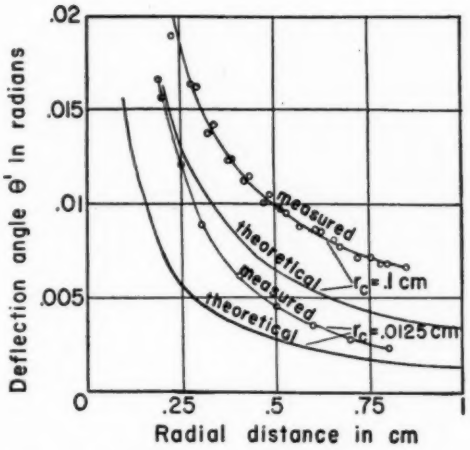


FIG. 5. Comparison of measured and computed deflection angles.

coordinate of the electron does not change during its traversal of the magnetron, and (2) that the electric field is constant in the axial direction. By proper choice of V_0 the first assumption was justified. The second assumption is invalid, since there are "fringe fields" around the ends of the magnetron. The effect of the electric fringe field can be determined directly by the electron optical shadow method.

III. MEASUREMENTS

3.1. Experimental Procedure

In order to measure the effect of the fringe field, the difference between the deflection angle measured electron-optically in the magnetron without space charge and the theoretically expected deflection angle according to Eqs. (1) and (2) is ascribed to the fringe field effect.

Figures 3 and 4 show the patterns observed on the fluorescent screen, Fig. 3 with the electric field and Fig. 4 without electric field, i.e., with the anode shorted to the cathode.

The X-shaped strips are the shadows of the cathode supporting bars. As these bars are aligned parallel to each other, the angle between their respective shadows is a measure of the strength of the homogeneous magnetic field. In Fig. 3 the thickness of these shadows is seen to vary. The shadow thickness of the supporting bars is roughly inversely proportional to the field strength E.

The small black spots in Figs. 3 and 4 are dots of Aquadag painted on each of the wire screens. They enable us to recognize the same wire shadow on the two pictures to be compared. From the geometric disposition of these "tags," it is possible to count correctly the number of wire shadows hidden by X-shaped strips separating the pattern into 4 individual sections.

The deflection angles, θ', computed for various radial distances from Figs. 3 and 4 according to the instructions given in the Appendix, are shown in Fig. 5. The

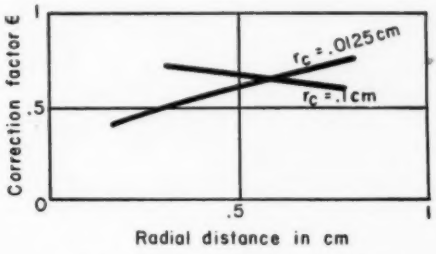


FIG. 6. Fringe field correction factor.

experiment was carried out for a magnetron without space charge and with two different radii for the central electrode (r_c=0.0125 cm and r_c=0.1 cm). The theoretically expected deflection angles for both cathode radii are given in the same Fig. 5. The precision of the measured points is seen to be better than ±5 percent. A systematic shift of the whole curve of ±10 percent is possible.

We now define a correction factor ϵ as the ratio of theoretically expected deflection angle to measured deflection angle at a given radius. The curves of ϵ *versus* r for both cathode radii are shown in Fig. 6.

Presence of space charge in the magnetron brings a new unknown factor into the field measurement. The fringe field of the tube with space charge is not the same as the fringe field of the tube in the absence of space charge. It has not been possible to compute quantitatively the fringe field effect of the magnetron with space charge. An estimate, based on qualitative arguments, shows that the error introduced in neglecting the effect of the space charge on the fringe field is small. Therefore, it has been assumed that the correction factors of Fig. 6 can be applied without significant error. A series of measurements of fields in magnetrons of different lengths or an experiment in which the viewing beam is passed transversely through the fringe field would allow us to compute the fringe field effect in the presence of space charge. The project was terminated, however, before these experiments could be performed.

3.2. Observation of the Magnetron

The investigation of the space charge has been carried out in the region of distinct cutoff, although a small anode excess current was still present. The degree of space charge saturation in the tube was varied during the experiments by changing the cathode heating current. The space charge patterns and the patterns of the magnetron without electric field necessary for the field computation were photographed.[8]

Two cathodes were used; one was a tungsten cathode of r_c=0.0125 cm, and the other was a tantalum tubing cathode of r_c=0.1 cm. Both types of cathodes were pulled straight by springs and aligned carefully. In spite of this care, a parallel displacement of the cathodes with

[8] For simplicity of presentation only a few experiments are discussed here.

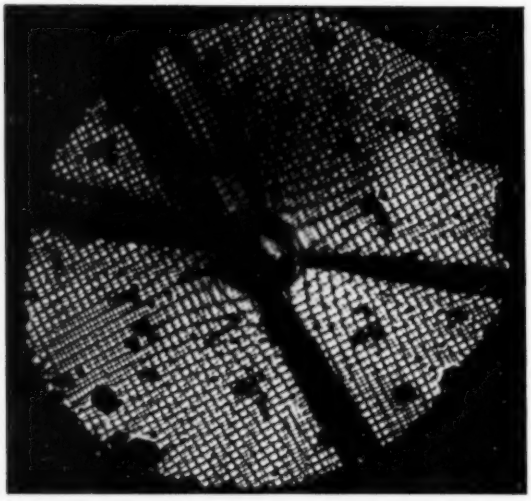


FIG. 7. Pattern obtained with space charge.

respect to the tube axis before and after the experiment was observed. The displacement was sometimes as large as 0.05 cm.

The typical pattern of a space charge saturated magnetron ($r_c=0.0125$ cm and $V_a=45$ volts) is shown in Fig. 7. The effect of the space charge is very marked. The $1/r$ field of the magnetron without space charge (compare with Fig. 3) is partially neutralized by the presence of the space charge so that the distortion of the pattern in Fig. 7 is less than that of Fig. 3.

For this pattern (Fig. 7) and other patterns, the deflection angles have been computed as a function of the radius. The fringe field correction has been applied and the electric field curves as a function of the radius for the various experiments have been obtained. Curves (a) and (b) in Fig. 8 are for $r_c=0.0125$ cm, curve (a) for a space charge saturated magnetron, curve (b) for a temperature-limited magnetron. Figure 9 shows the field distribution in a temperature-limited magnetron with a large cathode ($r_c=0.1$ cm).

The curves contain measured points between $r=0.3$ and $r=0.8$ cm and are extrapolated on the right to the anode and on the left to the cathode, with dotted lines. Between the measured part of the curves and the anode the field decreases approximately linearly (proportional to $1/r$) to a positive value at the anode. On the cathode side the general shape of the curve depends strongly on the degree of saturation of the tube. On curve (a) in Fig. 8, which represents the field distribution in a saturated magnetron, the field is approximately zero at the cathode. For a temperature-limited magnetron (Fig. 8(b) and Fig. 9) the field at the cathode can have any positive value, depending on the degree of space charge saturation. The unknown part of the curves is correct when the complete integral of the proposed field curve is equal to the potential difference between the anode and the cathode. The curves so adjusted are seen to be consistent with the space charge conditions of the test magnetron. The systematic shift of the whole curve due to the error of the fringe field is estimated to be 20 percent, twice as much as in the field measurement without space charge (compare Sec. 3.1).

The electric field distribution expected from the Hull-Brillouin theory is also indicated on Fig. 8. It can be compared only with experiment (a) in Fig. 8 because the theory assumes space charge saturation. The theoretical field is seen to rise linearly up to the cloud radius. From there it decreases like $1/r$. Both the experimental and the theoretical curves have a maximum, but the experimental curve does not conform to the shape of the theoretical curve. The degree of uncertainty in the fringe field correction is not sufficient to account for the difference.

The similarity of Fig. 9 and Fig. 8(b) indicates that the field distribution is modified only slightly, if at all, by the use of cathodes of different radii.

A very interesting phenomenon occurred unexpectedly during one of the measurements as the temperature of the cathode was increased to about the saturation point. A bright ring[9] on the screen located

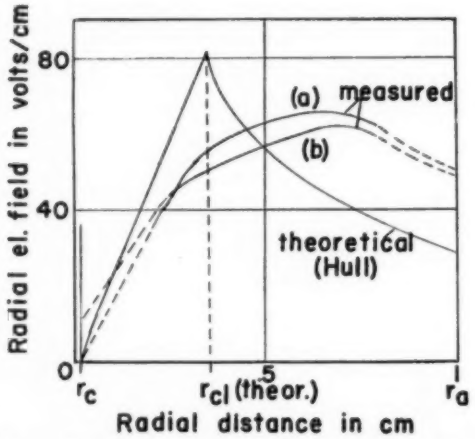


FIG. 8. Radial field distribution, $V_a=45$ volts, $r_c=0.0125$ cm, (a) space charge saturated magnetron, (b) temperature-limited magnetron.

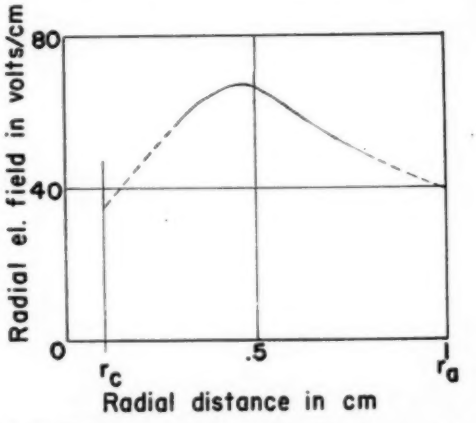


FIG. 9. Radial field distribution, $V_a=45$ volts, $r_c=0.1$ cm (temperature-limited magnetron).

[9] L. Marton and D. L. Reverdin, J. Appl. Phys. 21, 842 (1950).

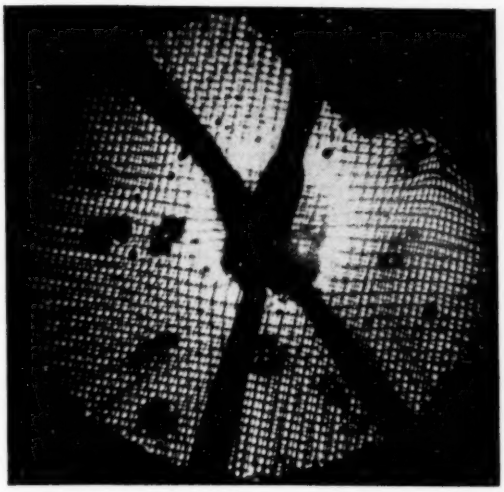


FIG. 10. Pattern showing ring phenomenon.

between the cathode and the anode appeared; this is plainly visible in Fig. 10. This pattern reveals a region within the ring with relatively small distortion. A thorough computation of this pattern cannot be performed, because the wire shadows inside the ring cannot be identified with sufficient certainty. The electric field in the magnetron, however, can be estimated (compare Sec. 3.1) by the thickness of the shadows of the cathode suspension bars (see Fig. 11).

An investigation indicated that the directly observed ring pattern was due to the effect of the electron space charge inside the magnetron and not to any additional effect inside or outside the magnetron. The ring would get weaker and finally disappear as the temperature of the cathode was lowered. It was very difficult to reproduce the effect in later experiments and the rings observed later were never as marked.

Another effect was observed when the filament of the magnetron was poorly aligned or intentionally inclined against the optical axis. In this case the pattern was not steady. A slow rotation of about one cycle per second was observed. The frequency of rotation could be slightly increased by raising the cathode temperature. The anode current oscillated with the same frequency. It appeared as if a portion of the space charge cloud was revolving around the cathode at a relatively slow rate.

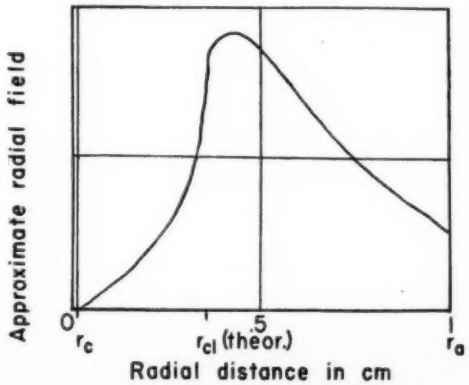


FIG. 11. Approximate field distribution for ring pattern.

IV. RESULTS AND DISCUSSION

4.1. Space Charge Distribution

Since the measurement of the electric field involves experimental error due principally to uncertainty in the fringe field, the resulting space charge distribution curves are of a qualitative nature. Curve (2) in Fig. 12 represents a typical space charge density distribution computed from the field curve in Fig. 8(a). The measured space charge distributions do not agree with the theoretical space charge distribution according to the Hull-Brillouin theory shown in curve (3) in Fig. 12. The measured curve tends to have a plateau near the cathode. The space charge density starts to decrease for a smaller radius than the theoretical cloud radius; the decrease of the density occurs gradually.

The space charge distribution for the ring picture (Fig. 10) can be given qualitatively from the photograph as shown in curve (1) in Fig. 12. The shape of the space charge distribution is seen to have a very marked maximum at a radius approximately equal to the theoretical radius.[10] The maximum is not predicted by the Hull theory. A detailed numerical computation by Page and Adams[11] has, however, resulted in a similar space charge distribution (see curve (4) in Fig. 12). Such a maximum may indicate that the electron stream which produces this distribution has a double stream character.

The experimental magnetron seemed more frequently to produce a space charge distribution which falls off gradually (curve (2)) instead of a space charge distribution with a distinct maximum near the cloud radius (curve (1)). The form of the space charge distribution must be closely related to the symmetry in the magnetron, because the symmetry is the only parameter which did vary in the experiments. Although the cathode was always very carefully aligned with respect to

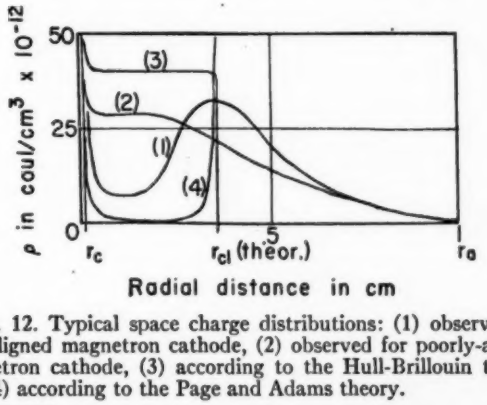


FIG. 12. Typical space charge distributions: (1) observed for well aligned magnetron cathode, (2) observed for poorly-aligned magnetron cathode, (3) according to the Hull-Brillouin theory, and (4) according to the Page and Adams theory.

[10] In reality this is a second maximum, as the theory predicts that the space charge density is maximum at the cathode.

[11] L. Page and N. J. Adams, Phys. Rev. 69, 494 (1946).

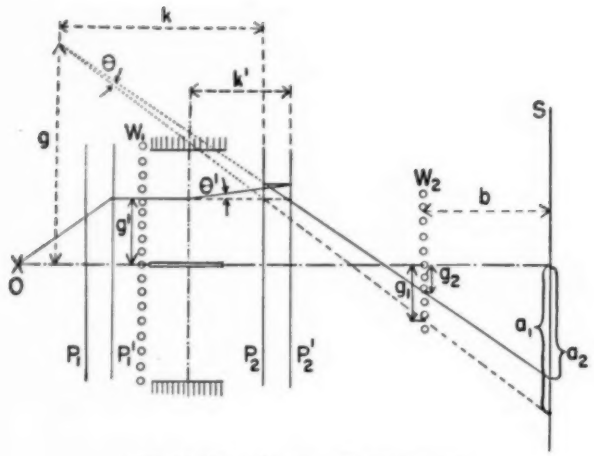


FIG. 13. Ray diagram of experiment.

the anode, the alignment could not be controlled after the cathode was heated. The probability of having a poorly aligned cathode is greater than that of having a well-aligned cathode. This might indicate that the space charge ring (Fig. 12 (1)) corresponds to a well-aligned magnetron and the falling off space charge distribution (Fig. 12 (2)) to a poorly aligned magnetron. A more thorough investigation would be very desirable here.

A tentative explanation for the mechanism which related the different shapes of observed space charge distributions, can be given on the basis of Haeff's theory of noise amplification.[12] The strong noise present in the dense magnetron space charge is visible in Figs. 7 and 10 by studying the sharpness of the shadows. The noise theory explains not only the presence of excess current but also the fact that all the observed space charge distributions fall off gradually toward the anode. If the cathode is not well aligned, the distribution curve (1) in Fig. 12 cannot exist, because of the asymmetric noise current to the anode. The presence of the space charge ring would then depend critically on the alignment. The observation of the rotation of the space charge in a magnetron, the cathode of which is intentionally inclined against the optical axis, supports this argument. Here, the field distribution along the magnetron axis is essentially asymmetric with respect to the midplane of the magnetron. A spiraling motion of the whole space charge cloud at a slow rate establishes itself. This observation indicates that space charge motions (as, for instance, spiraling of a cloud or transition from ring-shaped cloud to radially falling off cloud) induced by noise exist and occur slowly.

ACKNOWLEDGMENT

The author is indebted to Dr. L. Marton, Chief of the Electron Physics Section, under whose direction the work was done. The author gratefully acknowledges the assistance of Mr. John A. Simpson, Dr. R. T. Young, and Professor T. B. Brown, who were very helpful in discussing the experiments, and Mr. Simon H. Lachenbruch for the mathematical derivations.

[12] A. V. Haeff, Phys. Rev. 74, 1532 (1948).

APPENDIX. DERIVATION OF THE DEFLECTION FORMULA

The physical quantities necessary for the derivation of the deflection formula are defined in Fig. 13. The dashed line represents a trajectory with no electric field inside the object, and the solid line indicates the trajectory after it has been deflected by an angle θ' by the unknown electric field.

The image distance k is conjugate to the object distance k', g is conjugate to g', θ is conjugate to θ', the principal planes P_1 and P_2 are conjugate to P_1' and P_2'.

The derivation is based on the expression for the lateral image magnification:

$$g'/k' = g/k. \tag{4}$$

From Fig. 13, the identity

$$\theta' k' = \theta k, \tag{5}$$

can easily be seen.

Combining these two equations gives the deflection angle θ' expressed as a function of θ, g, and g' by the relation

$$\theta' = \theta g/g'. \tag{6}$$

The positions of the asymptotes are determined from the pattern on the fluorescent screen, the position of the second wire screen, and the number of meshes per unit length of a screen. From straightforward geometric considerations, one finds the height of the intersection of the asymptotes,

$$g = [(a_1-a_2)(a_1-g_1)/\{(a_1-a_2)-(g_1-g_2)\}]-a_1, \tag{7}$$

and the angle between the asymptotes,

$$\theta = [(a_1-a_2)-(g_1-g_2)]/b, \tag{8}$$

both as a function of the measurable quantities a_1, a_2, g_1, g_2, and the distance b.

Substituting Eqs. (7) and (8) into Eq. (6), the final expression for the radial deflection angle due to the field change turns out to be

$$\theta' = (a_2 g_1 - a_1 g_2)/bg'. \tag{9}$$

The following shall serve as instruction for the application of the deflection formula (9). From two photographs, for instance, Fig. 3 with electric field and Fig. 4 without electric field, g', a_1, a_2, g_1, and g_2 can be measured. The quantity g' is the independent parameter and is obtained by counting the number of wire shadows of the first wire screen from the center of symmetry along a chosen radial axis of measurement and multiplying it by the number of meshes per unit length of the screen. The distance from the center of symmetry to each one of the wire shadows g' is measured with a scale on each of the photographs; from Fig. 3 we obtain a_2, from Fig. 4 we get a_1. Then the fractional wire numbers of the shadows of the second wire screen at which the shadows of each of the chosen wire numbers g' of the first screen appear to lie are counted from the origin (center of symmetry); multiplied by the number of meshes per unit length of the second wire screen, this gives the quantities g_1 and g_2 for the second and first, respectively. These measurements are tabulated for each value of g' and the deflection angle θ is computed for each value of g' according to Eq. (10).[13]

[13] This set of measurements has to be carried out in a number of radial directions. By comparison of the plots of the computed deflection angles in the various directions, a poor choice of origin can be avoided.

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

A Note on Reflection and Transmission

BELA A. LENGYEL
Naval Research Laboratory, Washington, D. C.
(Received June 2, 1950)

The quantities characterizing wave propagation through the interface of two media or through a transmission line containing obstacles are the two reflection coefficients r_1 and r_2 and the two transmission coefficients t_1 and t_2. It is shown that for any lossless discontinuity $|r_1|=|r_2|$, $\arg t_1=\arg t_2$, and that $\arg r_1+\arg r_2-2\arg t_1=\pm\pi$. These relations are applied to the calculation of reflections from sheets of dielectric and metal-plate media, the latter being distinguished by the fact that $\arg t_1\neq 0$.

INTRODUCTION

REFLECTION and transmission at the surface of dielectric media is most frequently treated in terms of the wave theory and immediate use is made of the continuity of the electric vector at the boundary. Similar phenomena in transmission lines are usually discussed by the use of the impedance concept and equivalent circuits. Thinking of this latter type leads to relations between dominant transmission modes which are not continuous at the boundary. It is shown here that useful phase and amplitude relations applicable to phenomena of both types can be obtained extremely easily by means of the scattering matrix of J. Schwinger. The amplitude relations are well known, the phase relations are not.

The phenomena to be considered are characterized by the condition that there is a discontinuity or anomaly of wave propagation in a limited region, (usually plane) but far from that region; on either side there is, in the general case, one wave traveling toward and one away from the discontinuity. The scattering problem is solved when the relations between these four waves are known. In the case of a plane dielectric interface the directions of propagation in the two media may be different. The angles θ_1 and θ_2 included between the rays and the surface normal satisfy the laws of geometrical optics. In the case of a transmission line all waves are collinear. This distinction, however, is immaterial; it is necessary only to examine the complex amplitudes of the associated four waves.

THE SCATTERING MATRIX

Let $z=0$ be the locus of the discontinuity; the material constants and transmission line dimensions may be different on the two sides of this plane. We will speak of medium 1 for $z<0$ and medium 2 for $z>0$. Energy storage in the neighborhood of $z=0$ is permitted in the form of oscillations not associated with the transport of energy. The four dominant waves in the system may be extended mathematically to the origin as follows. When, for large negative z, a wave has the asymptotic form $b_1\exp j(\omega t+kz)$, the quantity b_1 will be called the amplitude of the wave at the origin.

Let a_1 and a_2 be complex numbers proportional to the electric amplitudes at $z=0$ of the waves traveling toward the discontinuity, and let b_1 and b_2 have a similar meaning for waves traveling away from the discontinuity. The subscripts are indicative of the media 1 (left) and 2 (right). Let $\mathbf{a}$ be a vector with components a_1 and a_2 and let $\mathbf{b}$ have a similar meaning. Because of the linear character of Maxwell's equations and the boundary conditions, the vector $\mathbf{b}$ depends on $\mathbf{a}$ linearly

$$\mathbf{b}=\mathbf{S}\mathbf{a}. \tag{1}$$

The scattering matrix $\mathbf{S}$ connects the incoming and the outgoing waves. Let the reflection coefficient for waves incident from the left be r_1, the transmission coefficient be t_1, and the corresponding quantities for waves incident from the right be r_2 and t_2; then

$$\mathbf{S}=\begin{pmatrix} r_1 & t_2 \\ t_1 & r_2 \end{pmatrix}. \tag{2}$$

A discontinuity, or junction, will be called lossless when it permits complete reversal of wave propagation. Such a reversal is equivalent to the interchange of the vectors $\mathbf{a}$ and $\mathbf{b}$ and the reversal of time, requiring the replacement of j by $-j$. The matrix $\mathbf{S}$ is then replaced by its complex conjugate $\mathbf{S}^*$ and the equation $\mathbf{a}=\mathbf{S}^*\mathbf{b}$ follows for any pair of vectors $\mathbf{a}$, $\mathbf{b}$ satisfying $\mathbf{b}=\mathbf{S}\mathbf{a}$. Therefore, for a lossless discontinuity

$$\mathbf{S}\mathbf{S}^*=\mathbf{I}, \tag{3}$$

where $\mathbf{I}$ is the identity matrix.[1]

The matrix equation (3) is equivalent to the following four scalar equations:

$$|r_1|^2+t_2t_1^*=1, \tag{4}$$

$$|r_2|^2+t_1t_2^*=1, \tag{5}$$

$$r_1t_2^*+t_2r_2^*=0, \tag{6}$$

$$t_1r_1^*+r_2t_1^*=0. \tag{7}$$

From Eqs. (4) and (5) it follows that $|r_1|^2-|r_2|^2=t_1t_2^*-t_1^*t_2=2\,\mathrm{Im}\,t_1t_2^*$. Therefore, $|r_1|=|r_2|$, and $t_1t_2^*$ is real. In fact, from Eq. (5) it also follows that: $t_1t_2^*=1-|r_2|^2\geqq 0$ because $|r_2|$ cannot exceed one. Therefore, $\arg t_1=\arg t_2$. Let $\rho=|r_1|$ and $\tau=(t_1t_2^*)^{\frac{1}{2}}$

[1] See Montgomery, Dicke, and Purcell, *Principles of Microwave Circuits* (McGraw-Hill Book Company, Inc., New York, 1948), p. 149.

$=(1-\rho^2)^{\frac{1}{2}}$. Thus, we may write

$$r_1=\rho e^{-j\rho'},\quad r_2=\rho e^{-j\rho''},\quad t_1=q\tau e^{-j\tau'},\quad t_2=q^{-1}\tau e^{-j\tau'},\quad (8)$$

where the positive factor q depends only on the normalization of the amplitude.[2] From Eq. (6) it then follows that

$$\exp j(-\rho'+\tau')=-\exp j(\rho''-\tau')=\exp j(\rho''-\tau'\pm\pi).$$

Hence,

$$\rho'+\rho''-2\tau'=\pm\pi. \quad (9)$$

This last relation[3] is of fundamental importance. In the case of a dielectric sheet in air $\rho'=\pi$, $\rho''=0$, and $\tau'=0$. In the case of metal-plate media ρ', ρ'', and τ' depend on the ratio of the plate spacing to the wavelength and on the angle of incidence.

SYMMETRY CONSIDERATIONS

When a discontinuity possesses a high degree of symmetry, this fact is reflected in the matrix **S**, provided that the system of reference is properly chosen. In the case of a dielectric, or metal-plate slab, or a symmetrical radome sandwich, the plane $z=0$ may be made to coincide with the plane of symmetry of the structure. This applies equally to a symmetric iris in a wave guide. A reversal of the positive direction of z then leaves the matrix **S** unchanged. This reversal is equivalent to an interchange of the indices 1 and 2. Therefore, $r_1=r_2$, $t_1=t_2$. From Eq. (9) it follows that

$$|\rho'-\tau'|=\pi/2; \quad (10)$$

the phases of the reflection and the transmission coefficients of a symmetric lossless discontinuity differ by 90°, provided that the point of reference is the center of symmetry.

A lossless symmetric discontinuity is electrically characterized by two quantities, preferably ρ and τ', since $\tau=(1-\rho^2)^{\frac{1}{2}}$ and τ' has to satisfy Eq. (10).

By examining the effect of coordinate transformation on the matrix **S**, one may show that a general lossless discontinuity surrounded by the same medium is equivalent electrically to a symmetric one, provided the system of reference is properly chosen.

APPLICATIONS

The reflection and transmission coefficients of dielectric sheets are well known. Expressions of the magnitude of the reflection coefficient and the phase of the transmission coefficient are of primary interest. They are calculated from the coefficients pertaining to a single interface taking into account the multiple reflections between the front and back surfaces. Unfortunately, early advantage is generally taken of the real character of t_1; and, therefore, the formulas available in the literature do not apply when $\tau'\neq 0$. This case is of interest in connection with structures of metal-plate media; it can be covered easily by rearranging the usual calculations.

The general expressions for the reflection and the transmission coefficients of a slab surrounded by the same medium are found in the following form:[4,5]

$$R=r_1+(r_2t_1t_2\exp-2j\delta_2)/(1-r_2^2\exp-2j\delta_2), \quad (11)$$

$$T_1=(t_1t_2\exp-j\delta_2)/(1-r_2^2\exp-2j\delta_2), \quad (12)$$

where δ_2 is the phase delay between the two intersections of a normal of the slab with the surfaces. Capital letters are used to distinguish the coefficients pretaining to the slab from those pertaining to an interface. The coefficient R as given by Eq. (11) refers to the front surface of the slab, not its plane of symmetry. The argument of T_1 gives the phase of the wave emerging at the back surface and *not* the change of phase produced by the introduction of the slab into the path of the radiation. The latter, T, is $T_1\exp j\delta_1$, where δ_1 is the phase delay through a slab of thickness d made of medium 1. When medium 1 is air, which is usually the case,

$$\delta_1=2\pi d\cos\theta_1/\lambda.$$

From Eq. (11) with the aid of Eqs. (8) and (9) we get

$$|R|=2\rho|\sin\Psi|/[(1-\rho^2)^2+4\rho^2\sin^2\Psi]^{\frac{1}{2}}, \quad (13)$$

and from Eq. (12)

$$\arg T=2\tau'-\delta_1+\delta_2+\arg\{1-\rho^2\exp-2j\Psi\}, \quad (14)$$

where $\Psi=\rho''+\delta_2$. This last equation might be simplified to

$$\arg T=\rho'+\Psi'-\pi-\delta_1, \quad (15)$$

where Ψ' is defined as follows:

$$\tan\Psi'=F\tan\Psi,\quad F=(1+\rho^2)/(1-\rho^2). \quad (16)$$

The form of Eq. (13) is not new. It occurs in the literature for the case when $\rho''=0$ or π; i.e., when $|\sin\Psi|=|\sin\delta_2|$. The introduction of the angle Ψ' is justified by the fact that the difference $\Psi'-\Psi$ is generally small; in fact,

$$\Psi'-\Psi\approx\tan(\Psi'-\Psi)=(F-1)/\cot\Psi+F\tan\Psi. \quad (17)$$

Equations (13) and (15) contain the quantities which are most readily accessible by experiment with the present microwave techniques.

[2] When the quantities a_1, a_2, b_1, and b_2 are so chosen that $\frac{1}{2}|a_1|^2$, $\frac{1}{2}|a_2|^2$, etc., represent the mean energy flows, the scattering matrix **S** is symmetrical; $q=1$.

[3] Equation (9) was derived by R. M. Redheffer for the case of a series of lossless dielectric sheets. M.I.T. Rad. Lab. Report 483–12 (1945), unpublished.

[4] C. G. Montgomery, *Technique of Microwave Measurements*, Rad. Lab. Ser. Vol. 11 (McGraw-Hill Book Company, Inc., New York, 1947), p. 564.

[5] Cady, Carelitz, and Turner, *Radar Scanners and Radomes*, Rad. Lab. Ser. Vol. 26 (McGraw-Hill Book Company, Inc., New York, 1948), p. 355.

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

Reflection and Transmission at the Surface of Metal-Plate Media

BELA A. LENGYEL
Naval Research Laboratory, Washington, D. C.
(Received June 23, 1950)

The theory of Carlson and Heins has been extended; an expression is found for the reflection coefficient applicable in the presence of a diffracted beam. Tables and graphs are included for the coefficients associated with electromagnetic phenomena at the surface of metal-plate media. The magnitude of the reflection coefficient was measured at normal incidence and the phase of the transmission coefficient for angles of incidence from 0 to 25°. A satisfactory agreement with theory is obtained.

INTRODUCTION

METAL–PLATE structures are often employed in microwave lens design. They consist of a set of parallel, equidistant, conducting plates so oriented that the electric vector of the incident waves is parallel to the plates. The problem of finding the phase and energy relations at the surface of metal-plate media is a great deal more complicated than in the case of dielectric media. The basic theoretical work in this field was done by Carlson and Heins,[1] who solved the field problem rigorously for the case when the plane of incidence is perpendicular to the metal plates. They obtained analytic expressions for the reflection and transmission coefficients which are valid when the incident radiation is subject to a number of restrictions. The principal restriction limits the angle of incidence below a critical value determined by the ratio of wavelength to plate separation. When the critical value is exceeded, the edges of the plates act as a grating and diffracted beams arise. This possibility was excluded by Carlson and Heins. By extending the theory, I have found expressions for the reflection and transmission coefficients in the presence of a diffracted beam and prepared tables and graphs to show the energy distribution between various beams.

The problem pertaining to waves whose plane of incidence is not perpendicular to the metal plates can be reduced to the case discussed earlier, provided that the electric vector is parallel to the metal plates. The reflection and transmission coefficients of waves so polarized that their magnetic vector is parallel to the metal plates are independent of the wavelength. This was also shown by Carlson and Heins.[2]

A considerable amount of work is required to bridge the gap between the mathematical theory of Carlson and Heins and the variables accessible to the experimenter. The theory predicts the reflection and the transmission of a plane electromagnetic wave on a single boundary surface separating free space and a metal-plate medium extending (in depth) to infinity. In the case of a medium of finite depth, the reflection from the "back" surface combines with that at the front surface. In an experimental test of the theory it is then necessary either to compute reflection and transmission coefficients for media of finite depth from the single surface coefficients or to eliminate back surface reflections by the use of absorbing materials. The first method was employed in the experimental work reported here. It requires the knowledge of the phase as well as the magnitude of the coefficients associated with a single surface transition. The expressions for phase relations derived by Carlson and Heins are of the form of infinite series; their numerical evaluation presented a number of difficulties. The technique of the computation is described in a Naval Research Laboratory report;[3] tables of the coefficients are included in this article.

The basic coefficients pertaining to a single air-metal-plate-medium interface having been obtained, it became necessary to consider the problem of composing reflections originating at the front and back surfaces of a slab, or medium of uniform depth. The unusual phase relations prevailing in the case of metal-plate media make certain considerations necessary which are not required in the case of ordinary dielectric materials.

The second part of this article is devoted to experiments performed on metal-plate slabs with the objective of testing the predictions of the theory. Reflection measurements included in this article are confined to the case of normal incidence. A coordinated experimental project was undertaken at Oregon State College under the direction of Professor J. J. Brady, who specialized on reflection measurements at oblique incidence. His results will be reported in another article.

PART I. THEORY

1. Terminology and Elementary Properties

A parallel-plate or metal-plate medium is a part of space containing a number of parallel, equidistant, conducting sheets. Such a structure is shown in Fig. 1 with the plates perpendicular to the plane of the paper. The edges of the plates lie in a plane which will be called the surface of the medium. The plates are assumed to be infinitely conducting; the parameters determining the electrical properties of the medium are the plate separation a, the plate thickness t, and the shape of the surface

[1] J. F. Carlson and A. E. Heins, Quart. Appl. Math. 4, 313–329 (1947).
[2] A. E. Heins, and J. F. Carlson, Quart. Appl. Math. 5, 82–88 (1947).
[3] B. A. Lengyel, "Physical Optics of Metal-Plate Media. I," NRL Report 3534, Washington (1949).

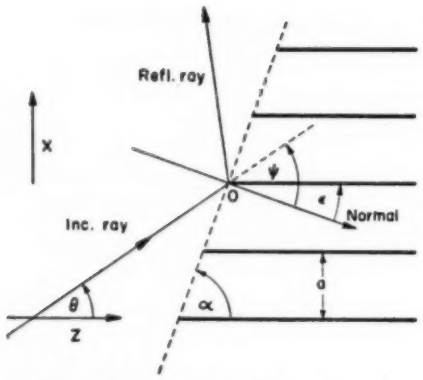


FIG. 1. Metal-plate medium (side view).

of the medium determined by the angle of stacking α. Propagation within the medium is determined by a, surface phenomena depend on α and t also. Other variables of interest are the plate spacing, $a'=a+t$, the stepback, $b=a'\cot\alpha$, and the angle $\epsilon=90°-\alpha$. It is sufficient to examine the case $0\leqq\epsilon<90°$.

Metal-plate media are generally intended for use in a well-defined frequency region and the parameters of the structure are so chosen that t is very small compared while the wavelength λ, a is of the same order as λ. The mathematical theory is based on the idealization $t=0$; $a'=a$.

We will consider only such an electromagnetic wave whose electric vector is parallel to the metal plates. In order to insure single-mode propagation between the plates we also assume that

$$1\leqq 2a/\lambda<2. \tag{1}$$

The ratio of the phase velocity in free space to the phase velocity in the medium is

$$n=[1-(\lambda/2a)^2]^{\frac{1}{2}}. \tag{2}$$

It is called the *index of refraction*. Let k be the propagation constant in free space, $k=2\pi/\lambda$; and let $\kappa=nk$ be the corresponding quantity in the medium. Equation (2) is equivalent to

$$k^2=\kappa^2+(\pi/a)^2. \tag{3}$$

The value of n determines the propagation within the plates and also the geometrical optics of the rays incident on the surface of the medium. Rays incident in a plane parallel to that of the plates are not constrained, while rays incident in all other planes are constrained by the plates. The unconstrained rays follow Snell's law. When rays are incident in the plane of maximum constraint, the (x, z) plane, the Poynting vector in the medium is always directed along the z-axis.

Following the example of Carlson and Heins we at first restrict our attention to rays incident in the (x, z) plane, and show later how the general case can be reduced to this one. A ray in the (x, z) plane may be characterized either by its elevation θ measured from the positive z axis, or by the angle ψ enclosed between the positive surface normal and the ray. Clearly, $\psi=\theta+\epsilon=90°-\alpha+\theta$ (Fig. 1). The variables ϵ and ψ are more convenient than the angles α and θ used by Carlson and Heins. For incident rays $\cos\psi>0$, while for reflected and diffracted rays $\cos\psi<0$.

When a plane wave is incident on the surface of a metal-plate medium, the magnitude of the electric field at the surface is not uniform, but a periodic function of position. Therefore, transmitted and reflected waves cannot be uniform plane waves as they are in the case of dielectrics. Only at a distance from the surface might one expect approximately uniform plane waves.

The edges of the metal plates act as a grating; and, consequently, for a single plane wave incident at an angle ψ, there will be reinforcement of scattered waves in all directions ψ_m which satisfy the relations: $\cos\psi_m<0$ and

$$\sin\psi_m-\sin\psi=m\lambda/a\sec\epsilon, \quad m=0, \pm1, \pm2, \cdots. \tag{4}$$

The ray reflected according to the law of geometrical optics corresponds to $m=0$. When $a\sec\epsilon>\lambda$, diffracted rays will arise for any angle of incidence. Since this situation is avoided in every application, we restrict ourselves to the case when

$$a\sec\epsilon/\lambda<1. \tag{5}$$

For $\epsilon\neq0$ this assumption is more restrictive than the condition of single mode propagation.[4] It eliminates values of m greater than one in magnitude from Eq. (4). In order that Eq. (4) be satisfied for $m=1$ or $m=-1$, $\sin\psi$ and $\sin\psi_m$ must be of opposite sign and $m\sin\psi_m>0$. Therefore, at most one diffracted beam will be present; it will lie on the same side of the normal as the incident beam.

When

$$1+|\sin\psi|<\lambda/a\sec\epsilon, \tag{6}$$

no diffracted beam is possible. This will be called Case A. In the opposite case there will be a diffracted beam, whose direction is specified by the angle ψ' determined from the equation

$$|\sin\psi-\sin\psi'|=\lambda/a\sec\epsilon. \quad (\cos\psi'<0.) \tag{7}$$

This will be called Case B.

2. The Theory of Rays in the Plane of Maximum Constraint[5]

When the orientation of the electric vector with respect to the plates is such as we have assumed, the electromagnetic problem is a scalar one, for $E_x=E_z=0$; and the magnetic components are derivable from

[4] Should we discard condition (5), we should encounter no difficulty in principle, but the discussion would become cumbersome, for it would be necessary to distinguish between a large number of cases.

[5] For a complete understanding of the next two sections, it is helpful to read the first article of Carlson and Heins.[1] In order to facilitate a comparative reading, the convention of using a time dependence in the form of $e^{-i\omega t}$ has been retained.

$\phi(x, z)=E_y$. The differential equation to be solved is

$$(\partial^2\phi/\partial x^2)+(\partial^2\phi/\partial z^2)+k^2\phi=0. \quad (8)$$

The boundary conditions are: $\phi=0$ on the plates, and an incident wave of the form $\exp ik(z\cos\theta+x\sin\theta)$. Carlson and Heins expressed this problem in the form of an integral equation and solved the equation by the use of the Fourier transform. The crucial step in the solution requires the splitting of a complex function $f(w)$ into two factors, one of which has no zeros or poles in a suitably defined upper half-plane, the other is similarly behaved in a lower half-plane, the two half-planes having a strip in common. The field is then obtained by evaluating a complex integral involving one of the factors.

We introduce a notation which is essentially identical to that of Carlson and Heins, the principal deviation being in the use of ψ and ϵ in place of the angles θ and α. Let

$$p(w)=a(k^2-w^2)^{\frac{1}{2}}, \quad (9)$$

$$q(w)=ak\sec\epsilon\sin\psi-wb, \quad (10)$$

$$f(w)=a\sin p/p(\cos p-\cos q), \quad (11)$$

$$I(w, x)=\{\sin[(x/a)-n'-1]p - e^{iq}\sin[(x/a)-n']p\}e^{iqn'}, \quad (12)$$

where n' is the greatest integer not exceeding x/a. The functions $K_+(w)$ and $K_-(w)$ are defined by the equation

$$f(w)=K_-(w)/K_+(w), \quad (13)$$

and the following conditions: K_+ and $1/K_+$ are analytic in the upper half-plane, while K_- and $1/K_-$ are analytic in the lower half-plane. The half-planes are separated by the curve C shown in Fig. 2.[6] These conditions determine K_+ and K_- only up to a common factor $\exp\chi(w)$, where $\chi(w)$ is an entire function. This function is chosen so as to insure convergence of the integrals, its form is immaterial for the present.

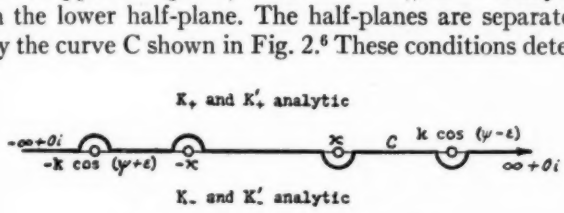


FIG. 2. The curve C in the plane w.

3. Determination of the Reflection Coefficients

Carlson and Heins found for the total electric field[7]

$$\phi(x, z)=\phi_{\text{inc}}+\frac{1}{2\pi i}\int_C \frac{K_+(w_1)I(w, x)e^{iwz}dw}{K_+(w)(w-w_1)\sin a(k^2-w^2)^{\frac{1}{2}}}, \quad (14)$$

where $w_1=k\cos(\psi-\epsilon)$. For negative z the path of integration C may be closed by a large arc in the lower half-plane, and for positive z by a similar path in the upper half-plane. The evaluation of the field then reduces to the evaluation of the residues. For the calculation of the far field only real poles need to be taken into account, complex poles leading to nonpropagating waves.

One of these poles is w_1, the others are the zeros of the function $K_+(w)\sin p(w)$, the roots of $p(w)=0$ being excepted because they also occur in the numerator. The function $K_+(w)\sin p(w)/p(w)$ has as its roots those roots of $\sin p(w)$ which lie in the upper half-plane, and in addition, those roots of the function $\cos p-\cos q$ which lie in the lower half-plane. Now $\sin p(w)=0$ and $p(w)\neq 0$ means

$$k^2-w^2=m^2\pi^2/a^2, \quad m=1, 2, \cdots. \quad (15)$$

Single-mode propagation between the plates requires that $\kappa=(k^2-\pi^2/a^2)^{\frac{1}{2}}$ is real, but that all solutions of Eq. (15) with $m>1$ are imaginary. Thus, $w=\kappa$ is the only real pole of the integrand (14) arising from $\sin p(w)=0$.

The remaining poles of the integrand are those solutions of the equation $\cos p=\cos q$ which lie in the lower half-plane. This transcendental equation is equivalent to the two sets of algebraic equations:

$$p-q=2\pi m, \quad (16)$$

$$p+q=2\pi m, \quad m=0, \pm 1, \pm 2, \cdots. \quad (17)$$

For $m=0$, Eq. (16) reduces to $p=q$; its solutions are $w_1=k\cos(\psi-\epsilon)$ and $w_2=-k\cos(\psi+\epsilon)$. (See Appendix A.) It is relatively easy to show that $w_3=-k\cos(\psi'+\epsilon)$ and $w_4=k\cos(\psi'-\epsilon)$ are solutions of $\cos p-\cos q=0$ if and only if

$$\sin\psi'-\sin\psi=m2\pi/ka\sec\epsilon=m\lambda/a\sec\epsilon, \quad (18)$$

for some integer m. (See Appendix A.) This is the elementary diffraction relation. In Case A a real solution is possible for $m=0$ only, while in Case B either $m=1$ or $m=-1$ is also permissible.

The pole $w_2=-k\cos(\psi+\epsilon)=k\cos(2\alpha-\theta)$ gives rise to a wave propagating in the direction which makes an angle $2\alpha-\theta$ with the positive z-axis or $180°-\psi$ with the positive surface normal. One may show that $|I(w_2, x)| = |\sin p(w_2)|$; therefore, if the complex amplitude of the reflected wave is r_1, then

$$\rho=|r_1|=|K_+(w_1)/K_+'(w_2)(w_2-w_1)|. \quad (19)$$

The pole $w_1=k\cos(\psi-\epsilon)=k\cos\theta$ gives rise to a wave which propagates in the direction of the incident wave and cancels the field of the latter within the metal-plate medium. The pole κ generates the transmitted wave whose z dependence is of the form $\exp i\kappa z$. The amplitude

[6] The curve C is essentially identical with the real axis, except in the vicinity of real poles and zeros of $f(w)$. See Appendix B.

[7] Equation (14) follows from formula (4.1) of Carlson and Heins (p. 325) when a few typographical errors are corrected and our abbreviations (9), (10), and (12) are introduced. The z-coordinate as used here is not measured from any fixed point in space, but essentially from the metal-plate surface. This is the consequence of a transformation $z_{\text{new}}=z_{\text{old}}+n'b$, introduced by Carlson and Heins on p. 318. Unfortunately the new variables are not distinguished from the old ones and one must remember that the origin is to be shifted on passing from a channel specified by $n'a\leqq x<(n'+1)a$ to the next.

of this wave is

$$(\pi/a^2\kappa)|[K_+(w_1)/(\kappa-w_1)K_+(\kappa)][1+\exp iq(\kappa)]| \times\sin(\pi x/a).$$

Case B differs from Case A in two respects: first, there is an additional ray present corresponding to w_4. Its direction is specified by the angle ψ' with the positive normal, its complex amplitude is $r^{(1)}$, where

$$\rho^{(1)}=|r^{(1)}|=|K_+(w_1)/K_+'(w_4)(w_4-w_1)|. \quad (20)$$

Second, ρ is a different function of k and ψ, because the functions $K_+(w)$ and $K_+'(w)$ are different in Case B.

In Case A, $f(w)$ is of the form

$$f(w)=(w-\kappa)(w+\kappa)G(w)/(w-w_1)(w-w_2)H(w), \quad (21)$$

where $G(w)$ and $H(w)$ are entire functions with no real zeros. Their complex zeros are all simple; they are located symmetrically about the real axis (see Appendix A). The Weierstrass factorization theorem then insures that G and H may be written in the following form:

$$G=G_1(w)G_2(w)e^{2g(w)}; \quad (22)$$

$$H=H_1(w)H_2(w)e^{2h(w)}, \quad (23)$$

where G_1 and H_1 are entire functions with all their roots in the upper half-plane, G_2 and H_2 are mirror images about the real axis of G_1 and H_1, respectively, g and h are entire functions. The definitions of K_+ and K_- then lead to

$$K_-=A\frac{w-\kappa}{w-w_1}\frac{G_1}{H_1}e^{g-h}e^{\chi(w)}, \quad (24)$$

$$K_+=A\frac{w-w_2}{w+\kappa}\frac{H_2}{G_2}e^{h-g}e^{\chi(w)}, \quad (25)$$

where A is an arbitrary constant, say, 1, and $\chi(w)$ is a convergence-producing entire function, shown by Carlson and Heins to be pure imaginary for real w.

Then, for any *real* w other than w_1 and $-\kappa$

$$|K_+K_-|=|fK_+^2|=\left|\frac{w-\kappa}{w+\kappa}\frac{w-w_2}{w-w_1}\right|, \quad (26)$$

since on the real axis $|G_1/G_2|=|H_1/H_2|=|e^{\chi}|=1$. Therefore,

$$|K_+(w_1)|^2=\left|\frac{w_1-\kappa}{w_1+\kappa}(w_1-w_2)\right|L_1, \quad (27)$$

where

$$L_1^{-1}=|\lim_{w\to w_1} f(w)(w-w_1)|,$$

and

$$|K_+'(w_2)|^2=\lim_{w\to w_2}\left|\frac{K_+(w)}{w-w_2}\right|^2=\left|\frac{w_2-\kappa}{(w_2+\kappa)(w_2-w_1)}\right|L_2, \quad (28)$$

where

$$L_2^{-1}=|\lim_{w\to w_2} f(w)(w-w_2)|.$$

The limits L_1 and L_2 are of equal modulus. Therefore, when the values of $|K_+|$ and $|K_+'|$ are substituted into Eq. (19),

$$\rho^2=\left|\frac{w_1-\kappa}{w_1+\kappa}\frac{w_2+\kappa}{w_2-\kappa}\right| \quad (29)$$

results. For the special case $\epsilon=0$,

$$\rho=\frac{k\cos\theta-\kappa}{k\cos\theta+\kappa}. \quad (30)$$

In Case B we replace Eq. (21) by

$$f(w)=\frac{(w-\kappa)(w+\kappa)G(w)}{(w-w_1)(w-w_2)(w-w_3)(w-w_4)H^*(w)}, \quad (31)$$

where H^* has properties similar to those of H in the previous case. The reasoning employed earlier leads to

$$|fK_+^2|=\left|\frac{w-\kappa}{w+\kappa}\frac{w-w_2}{w-w_1}\frac{w-w_4}{w-w_3}\right| \quad (32)$$

for all real values of w, except w_1, w_3, and $-\kappa$.[8] It follows then that the values of $|K_+(w_1)|^2$ and $|K_+'(w_2)|^2$ as computed from Eqs. (27) and (28) need to be multiplied by the factors $|(w_1-w_4)/w_1-w_3|$ and $|(w_2-w_4)/w_2-w_3|$, respectively, in order for values applicable to Case B to be obtained. Therefore,

$$\rho^2=\left|\frac{w_1-\kappa}{w_1+\kappa}\frac{w_2+\kappa}{w_2-\kappa}\frac{w_1-w_4}{w_1-w_3}\frac{w_2-w_3}{w_2-w_4}\right|. \quad (33)$$

The calculation of $\rho^{(1)}$ from Eq. (20) requires the knowledge of $|K_+'(w_4)|$. This is obtained from Eq. (32). In fact,

$$|K_+'(w_4)|^2=\left|\frac{w_4-\kappa}{w_4+\kappa}\frac{w_4-w_2}{w_4-w_1}\frac{L_4}{w_4-w_3}\right|, \quad (34)$$

where

$$L_4^{-1}=|\lim_{w\to w_4} f(w)(w-w_4)|.$$

It is easy to show that $L_1/L_4=\cos\psi/|\cos\psi'|$; therefore,

$$|r^{(1)}|^2=\left|\frac{w_1-\kappa}{w_1+\kappa}\frac{w_4+\kappa}{w_4-\kappa}\frac{w_4-w_3}{w_1-w_3}\frac{w_1-w_2}{w_4-w_2}\frac{\cos\psi}{\cos\psi'}\right|. \quad (35)$$

The quantity ρ^2 is the fraction of the incident energy thrown into the reflected beam. The proper measure of the energy in the diffracted beam is $-|r^{(1)}|^2\cos\psi'/\cos\psi$.

4. Numerical Results

For the special case $\epsilon=0$ numerical and graphical data are available concerning the coefficients discussed. Figure 3 shows the distribution of power among the various beams as a function of the angle of incidence for four selected values of n. Figure 4 shows ρ as a function of the angle of incidence for eight values of n. Detailed four-place tables of ρ *vs* ψ and ψ' *vs* ψ are available in NRL Report 3534.

[8] The curve C is so drawn that w_1, w_3, and κ are in the upper half-plane, while w_2 and w_4 are in the lower. See Appendix B.

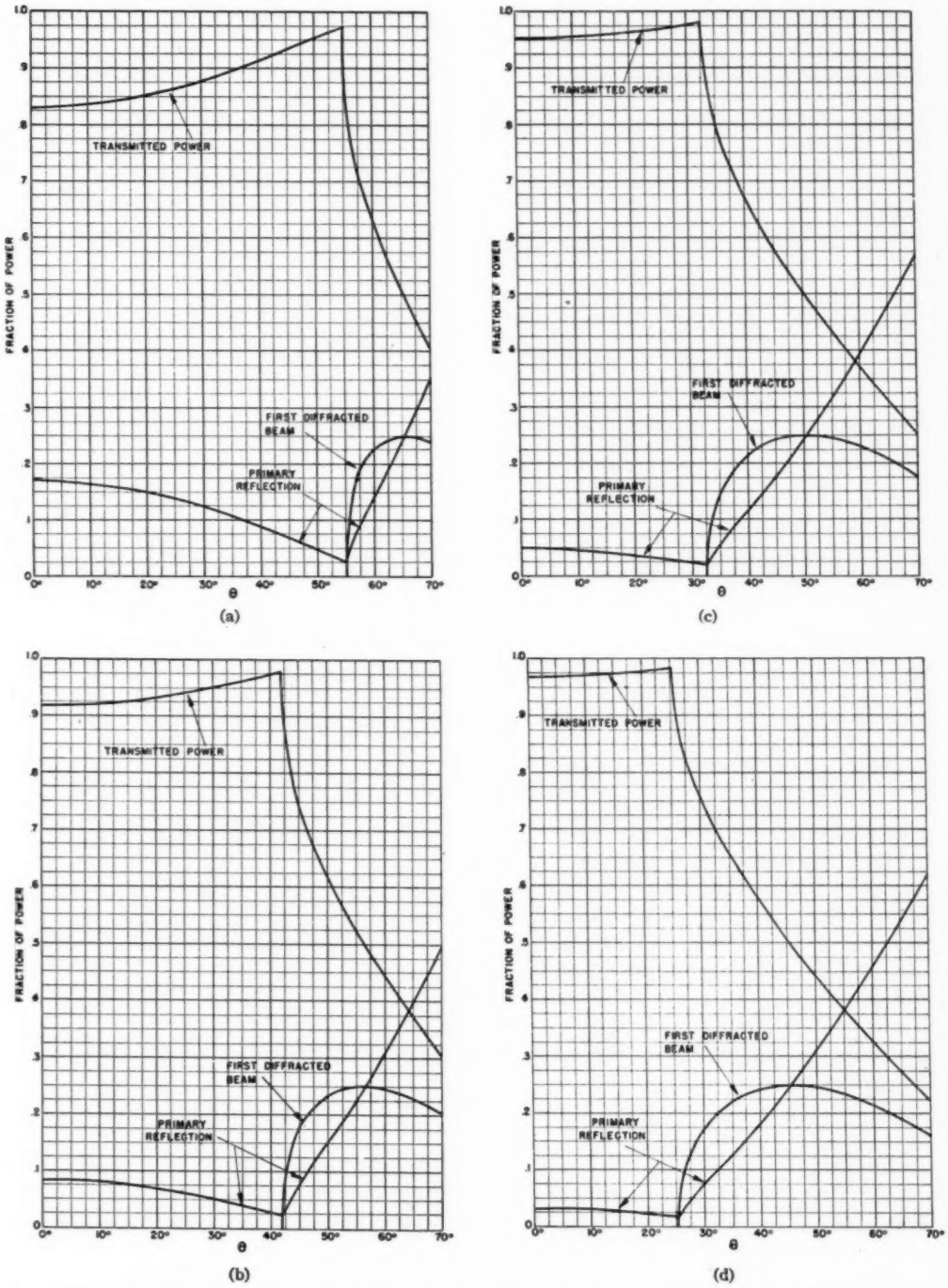


FIG. 3. Power distribution as a function of angle of incidence. (a) $2a/\lambda=1.10$, $n=0.417$. (b) $2a/\lambda=1.20$, $n=0.553$. (c) $2a/\lambda=1.30$, $n=0.639$. (d) $2a/\lambda=1.40$, $n=0.700$.

When there is no diffracted beam present, the computation of the reflection coefficient for the case $\epsilon\neq 0$ can be reduced to the case $\epsilon=0$. In fact, when the values of w_1 and w_2 are inserted and the index of refraction $n=\kappa/k$ is introduced, Eq. (29) takes the form,

$$\rho^2=\frac{\cos(\psi-\epsilon)-n}{\cos(\psi-\epsilon)+n}\,\frac{\cos(\psi+\epsilon)-n}{\cos(\psi+\epsilon)+n}. \tag{36}$$

We write $\rho=\rho(\psi, \epsilon, n)$, then

$$[\rho(\psi, \epsilon, n)]^2=\rho(\psi+\epsilon, 0, n)\rho(\psi-\epsilon, 0, n). \tag{37}$$

Thus, in Case A it is sufficient to tabulate ρ for $\epsilon=0$, because for $\epsilon\neq 0$, ρ may be obtained by combining two entries of this table. Such simple useful relationship does not exist in Case B. The expression corresponding

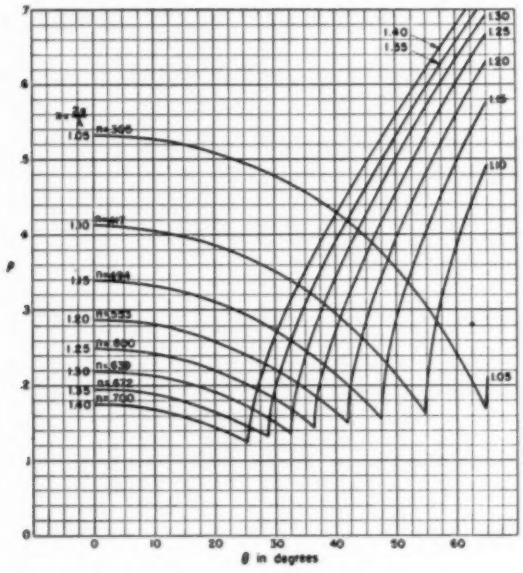


FIG. 4. Magnitude of the reflection coefficient for the case $\epsilon=0$ as a function of the angle of incidence.

to Eq. (36) in this case is

$$\rho^2=\left|\frac{\cos(\psi-\epsilon)-n}{\cos(\psi-\epsilon)+n}\,\frac{\cos(\psi+\epsilon)-n}{\cos(\psi+\epsilon)+n}\,\frac{\cos(\psi-\epsilon)+\cos(\psi'+\epsilon)}{\cos(\psi-\epsilon)-\cos(\psi'-\epsilon)}\times\frac{\cos(\psi+\epsilon)+\cos(\psi'-\epsilon)}{\cos(\psi+\epsilon)-\cos(\psi'+\epsilon)}\right|. \quad (38)$$

The computation of the phases of the reflection coefficients requires a great deal of effort even in the simplest case when no diffracted beams are present and when $\epsilon=0$. I did not attempt computation in any other case. In the absence of diffracted beams the following coefficients are of interest in connection with the transition of waves through a metal-plate-medium surface: the reflection coefficient r_1, and the transmission coefficient t_1 pertaining to waves incident from the air, and the corresponding quantities r_2 and t_2 pertaining to waves traveling in the opposite direction.

In the preceding article[9] I have shown that $|r_1|=|r_2|$, $\arg t_1=\arg t_2$ and that $\arg r_1+\arg r_2-2\arg t_1=\pm\pi$. Therefore, in addition to $\rho=|r_1|$, the phases $\rho'=\arg r_1$ and $\rho''=\arg r_2$ need to be computed. The function $\tau'=\arg t_1$ will then also be known. Figure 5 shows the phase functions for normal incidence, Tables I and II give $\rho'/2\pi$ and $\rho''/2\pi$ in a numerical form as functions of $2a/\lambda$ in a limited range which is of most interest from the point of view of applications.

5. Rays Incident in an Arbitrary Plane

The general ray considered in this section is subject to the requirement that its electric vector be parallel to the plates. Its plane of incidence, however, is arbitrary. Heins[10] and Brown[11] have shown that the problem associated with these rays can be reduced to that already solved by Carlson and Heins.

The process of reduction may be sketched as follows. Let the direction of the incident ray OP be characterized by the angles it includes with the coordinate axes, θ_1, θ_2, and θ_3, while the orientation of the metal-plate medium is identical to the one assumed in the earlier sections. The phase fronts of the incident waves are of the form

$$x\cos\theta_1+y\cos\theta_2+z\cos\theta_3=\text{const.}$$

Clearly, the y-dependence of the fields and currents in the metal-plate medium must be the same as that of the incident plane wave. Therefore, differentiation with respect to y may be replaced by a multiplication by $ik\cos\theta_2$. Moreover, $E_x=0$ everywhere. The electromagnetic problem again becomes a scalar one; all field components are expressible in terms of E_z. In fact, we may write $E_z=V=\phi(x,z)\exp iky\cos\theta_2$; then

$$E_y=(i/k\cos\theta_2)\partial V/\partial z,$$

and the components of H are also expressible in terms of V and its derivatives. Maxwell's equations lead to the following wave equation,

$$(\partial^2\phi/\partial x^2)+(\partial^2\phi/\partial z^2)+k^2\sin^2\theta_2\phi=0. \quad (39)$$

The boundary conditions are that ϕ should vanish on

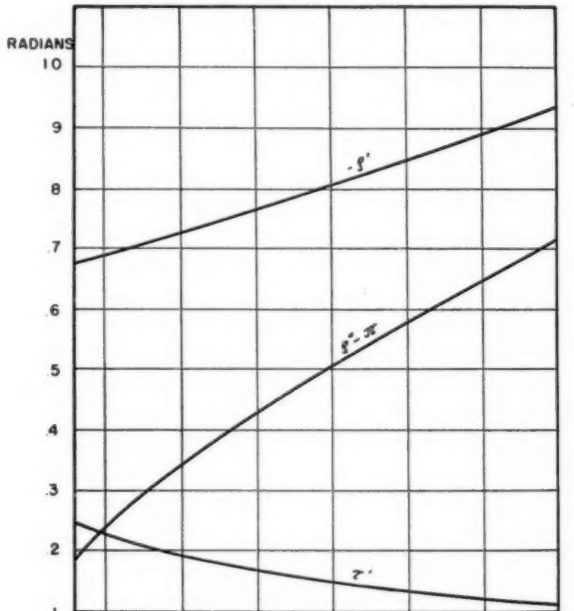


FIG. 5. The phase functions for normal incidence ($\epsilon=0$).

[9] B. A. Lengyel, J. Appl. Phys. 22, 263 (1951).

[10] A. E. Heins, oral communication, URSI Mtg., Washington, November, 1949.

[11] J. Brown, British Ministry of Supply rept., unpublished.

TABLE I. Tables of $-\rho'/2\pi$.

$x=\frac{2a}{\lambda}$	Angle of incidence 0°	5°	10°	15°	20°	25°
1.01	0.10487	0.10554	0.10755	0.11087	0.11544	0.12135
1.02	0.10605	0.10676	0.10888	0.11238	0.11723	0.12352
1.03	0.10724	0.10798	0.11021	0.11390	0.11904	0.12573
1.04	0.10844	0.10921	0.11155	0.11544	0.12088	0.12799
1.05	0.10964	0.11045	0.11290	0.11700	0.12274	0.13029
1.06	0.11084	0.11169	0.11427	0.11857	0.12464	0.13264
1.07	0.11205	0.11294	0.11565	0.12017	0.12657	0.13503
1.08	0.11327	0.11421	0.11704	0.12179	0.12853	0.13747
1.09	0.11450	0.11548	0.11844	0.12343	0.13053	0.13995
1.10	0.11573	0.11676	0.11985	0.12509	0.13256	0.14247
1.11	0.11697	0.11804	0.12128	0.12677	0.13464	0.14515
1.12	0.11821	0.11933	0.12272	0.12848	0.13675	0.14789
1.13	0.11946	0.12062	0.12418	0.13021	0.13891	0.15070
1.14	0.12071	0.12192	0.12565	0.13197	0.14112	0.15358
1.15	0.12197	0.12324	0.12713	0.13375	0.14338	0.15655
1.16	0.12324	0.12457	0.12863	0.13557	0.14568	0.15962
1.17	0.12452	0.12590	0.13015	0.13741	0.14804	0.16278
1.18	0.12581	0.12725	0.13168	0.13928	0.15046	0.16605
1.19	0.12710	0.12861	0.13323	0.14118	0.15294	0.16944
1.20	0.12840	0.12998	0.13480	0.14312	0.15548	0.17296
1.21	0.12971	0.13135	0.13639	0.14509	0.15809	0.1766
1.22	0.13103	0.13274	0.13799	0.14710	0.16076	0.1804
1.23	0.13235	0.13414	0.13961	0.14914	0.16351	0.1843
1.24	0.13368	0.13554	0.14125	0.15122	0.16633	0.1884
1.25	0.13502	0.13696	0.14292	0.15334	0.16924	0.1927
1.26	0.13637	0.13839	0.14461	0.15550	0.17224	0.1972
1.27	0.13773	0.13983	0.14631	0.15772	0.17533	0.2019
1.28	0.13910	0.14128	0.14804	0.15998	0.17852	0.2069
1.29	0.14048	0.14275	0.14979	0.16229	0.18183	0.2121
1.30	0.14187	0.14424	0.15157	0.16466	0.18526	0.2177
1.31	0.14327	0.14574	0.15339	0.16708	0.18882	0.2238
1.32	0.14468	0.14725	0.15523	0.16956	0.19253	0.2302
1.33	0.14610	0.14877	0.15709	0.17211	0.19638	0.2372
1.34	0.14753	0.15031	0.15898	0.17472	0.20040	0.2446
1.35	0.14898	0.15187	0.16091	0.17740	0.20459	0.2526

the plates and that for the incident wave,

$$\phi=\exp ik(x\cos\theta_1+z\cos\theta_3)$$
$$=\exp ik\sin\theta_2(x\sin\beta+z\cos\beta), \quad (40)$$

where β is the angle included between the z-axis and the projection of OP on the xz-plane. The differential equation and the boundary conditions are identical with those described in Sec. 2, except for the fact that in the present case $k\sin\theta_2$ takes the place of k and β takes the place of the angle θ. The formulas of Carlson and Heins may then be used in the general case, provided those substitutions are made.

It is interesting to calculate the magnitude of the reflection coefficient of unconstrained rays for the special case $\epsilon=0$. These rays are incident in the yz-plane; therefore, $\beta=0$. Their angle of incidence is $\psi=\theta_3=90°-\theta_2$. In Eq. (30) we replace θ by 0, k by $k'=k\sin\theta_2=k\cos\psi$, and κ by κ', where κ' is associated with k' in the same way as κ is associated with k; i.e., $k'^2=\kappa'^2+\pi^2/a^2$. It follows that

$$k^2\sin^2\psi=\kappa^2-\kappa'^2. \quad (41)$$

Let ψ' be the angle of refraction associated with the angle of incidence ψ according to Snell's law; i.e.,

$$n=\kappa/k=\sin\psi/\sin\psi'. \quad (42)$$

From Eqs. (41) and (42) then follows $\kappa'=\kappa\cos\psi'$, and

$$\rho=\frac{k'-\kappa'}{k'+\kappa'}=\frac{k\cos\psi-\kappa\cos\psi'}{k\cos\psi+\kappa\cos\psi'}. \quad (43)$$

Repeated use of Eq. (42) leads to

$$\rho=\sin(\psi'-\psi)/\sin(\psi'+\psi). \quad (44)$$

which is identical with one of Fresnel's formulas.

PART II. EXPERIMENTS

The experiments described here are a part of a program undertaken with the purpose of testing the predictions of the theory developed in Part I. In this article we restrict ourselves to the results of experimentation at the Naval Research Laboratory.[12] These include measurements of the magnitude of the reflection coefficient at normal incidence and measurements of the phase of the transmission coefficient with the angle of incidence varying from 0° to 25°.

TABLE II. Tables of $\rho''/2\pi$.

$x=\frac{2a}{\lambda}$	Angle of incidence 0°	5°	10°	15°	20°	25°
1.01	0.51670	0.51693	0.51761	0.51877	0.52044	0.52269
1.02	0.52371	0.52404	0.52503	0.52671	0.52915	0.53244
1.03	0.52916	0.52958	0.53082	0.53295	0.53605	0.54022
1.04	0.53381	0.53430	0.53578	0.53832	0.54200	0.54699
1.05	0.53794	0.53851	0.54021	0.54313	0.54738	0.55315
1.06	0.54173	0.54236	0.54428	0.54758	0.55239	0.55895
1.07	0.54525	0.54596	0.54810	0.55177	0.55713	0.56448
1.08	0.54857	0.54935	0.55171	0.55575	0.56167	0.56981
1.09	0.55173	0.55258	0.55514	0.55957	0.56605	0.57500
1.10	0.55475	0.55567	0.55845	0.56326	0.57032	0.58009
1.11	0.55766	0.55865	0.56165	0.56684	0.57450	0.58512
1.12	0.56047	0.56153	0.56476	0.57034	0.57860	0.59011
1.13	0.56319	0.56433	0.56778	0.57378	0.58266	0.59510
1.14	0.56585	0.56706	0.57075	0.57716	0.58669	0.60010
1.15	0.56844	0.56973	0.57366	0.58050	0.59071	0.60512
1.16	0.57098	0.57235	0.57653	0.58381	0.59471	0.61019
1.17	0.57347	0.57492	0.57936	0.58709	0.59871	0.61531
1.18	0.57592	0.57745	0.58215	0.59035	0.60271	0.62050
1.19	0.57833	0.57994	0.58491	0.59360	0.60675	0.62577
1.20	0.58070	0.58240	0.58765	0.59684	0.61083	0.63115
1.21	0.58304	0.58484	0.59037	0.60009	0.61494	0.63666
1.22	0.58536	0.58726	0.59309	0.60335	0.61909	0.64230
1.23	0.58766	0.58964	0.59579	0.60661	0.62329	0.64809
1.24	0.58993	0.59202	0.59848	0.60989	0.62756	0.65406
1.25	0.59218	0.59438	0.60117	0.61319	0.63191	0.66024
1.26	0.59442	0.59672	0.60384	0.61652	0.63634	0.66665
1.27	0.59664	0.59905	0.60652	0.61988	0.64086	0.67329
1.28	0.59884	0.60137	0.60921	0.62326	0.64547	0.68022
1.29	0.60104	0.60369	0.61190	0.62669	0.65019	0.68749
1.30	0.60323	0.60599	0.61460	0.63016	0.65506	0.69514

[12] The experimental program is described in greater detail in "Physical Optics of Metal Plate Media. II," NRL Report 3599, Washington (1950).

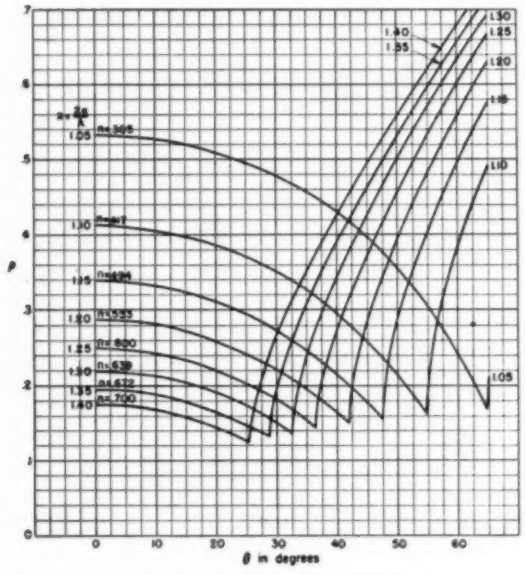


FIG. 4. Magnitude of the reflection coefficient for the case $\epsilon=0$ as a function of the angle of incidence.

to Eq. (36) in this case is

$$\rho^2=\left|\frac{\cos(\psi-\epsilon)-n}{\cos(\psi-\epsilon)+n}\,\frac{\cos(\psi+\epsilon)-n}{\cos(\psi+\epsilon)+n}\,\frac{\cos(\psi-\epsilon)+\cos(\psi'+\epsilon)}{\cos(\psi-\epsilon)-\cos(\psi'-\epsilon)}\times\frac{\cos(\psi+\epsilon)+\cos(\psi'-\epsilon)}{\cos(\psi+\epsilon)-\cos(\psi'+\epsilon)}\right|. \quad (38)$$

The computation of the phases of the reflection coefficients requires a great deal of effort even in the simplest case when no diffracted beams are present and when $\epsilon=0$. I did not attempt computation in any other case. In the absence of diffracted beams the following coefficients are of interest in connection with the transition of waves through a metal-plate-medium surface: the reflection coefficient r_1, and the transmission coefficient l_1 pertaining to waves incident from the air, and the corresponding quantities r_2 and l_2 pertaining to waves traveling in the opposite direction.

In the preceding article[9] I have shown that $|r_1|=|r_2|$, $\arg l_1=\arg l_2$ and that $\arg r_1+\arg r_2-2\arg l_1=\pm\pi$. Therefore, in addition to $\rho=|r_1|$, the phases $\rho'=\arg r_1$ and $\rho''=\arg r_2$ need to be computed. The function $\tau'=\arg l_1$ will then also be known. Figure 5 shows the phase functions for normal incidence, Tables I and II give $\rho'/2\pi$ and $\rho''/2\pi$ in a numerical form as functions of $2a/\lambda$ in a limited range which is of most interest from the point of view of applications.

5. Rays Incident in an Arbitrary Plane

The general ray considered in this section is subject to the requirement that its electric vector be parallel to the plates. Its plane of incidence, however, is arbitrary.

[9] B. A. Lengyel, J. Appl. Phys. 22, 263 (1951).

Heins[10] and Brown[11] have shown that the problem associated with these rays can be reduced to that already solved by Carlson and Heins.

The process of reduction may be sketched as follows. Let the direction of the incident ray OP be characterized by the angles it includes with the coordinate axes, θ_1, θ_2, and θ_3, while the orientation of the metal-plate medium is identical to the one assumed in the earlier sections. The phase fronts of the incident waves are of the form

$$x\cos\theta_1+y\cos\theta_2+z\cos\theta_3=\text{const.}$$

Clearly, the y-dependence of the fields and currents in the metal-plate medium must be the same as that of the incident plane wave. Therefore, differentiation with respect to y may be replaced by a multiplication by $ik\cos\theta_2$. Moreover, $E_x=0$ everywhere. The electromagnetic problem again becomes a scalar one; all field components are expressible in terms of E_z. In fact, we may write $E_z=V=\phi(x,z)\exp iky\cos\theta_2$; then

$$E_y=(i/k\cos\theta_2)\partial V/\partial z,$$

and the components of H are also expressible in terms of V and its derivatives. Maxwell's equations lead to the following wave equation,

$$(\partial^2\phi/\partial x^2)+(\partial^2\phi/\partial z^2)+k^2\sin^2\theta_2\phi=0. \quad (39)$$

The boundary conditions are that ϕ should vanish on

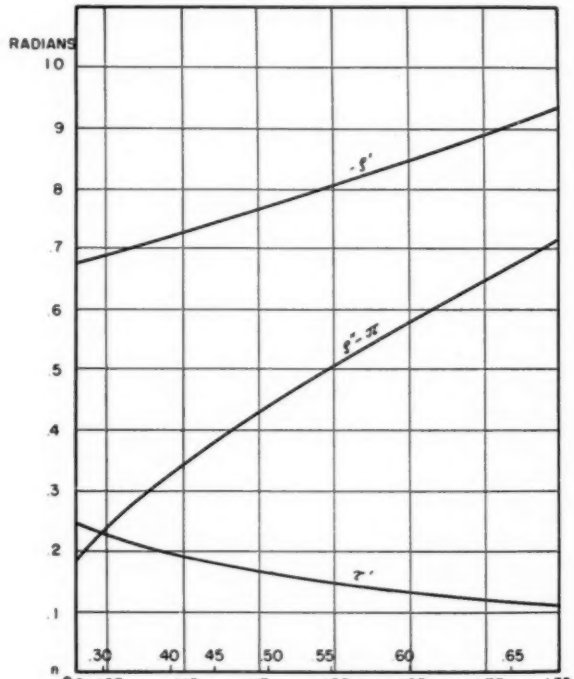


FIG. 5. The phase functions for normal incidence ($\epsilon=0$).

[10] A. E. Heins, oral communication, URSI Mtg., Washington, November, 1949.

[11] J. Brown, British Ministry of Supply rept., unpublished.

TABLE I. Tables of $-\rho'/2\pi$.

$x=\frac{2a}{\lambda}$	Angle of incidence 0°	5°	10°	15°	20°	25°
1.01	0.10487	0.10554	0.10755	0.11087	0.11544	0.12135
1.02	0.10605	0.10676	0.10888	0.11238	0.11723	0.12352
1.03	0.10724	0.10798	0.11021	0.11390	0.11904	0.12573
1.04	0.10844	0.10921	0.11155	0.11544	0.12088	0.12799
1.05	0.10964	0.11045	0.11290	0.11700	0.12274	0.13029
1.06	0.11084	0.11169	0.11427	0.11857	0.12464	0.13264
1.07	0.11205	0.11294	0.11565	0.12017	0.12657	0.13503
1.08	0.11327	0.11421	0.11704	0.12179	0.12853	0.13747
1.09	0.11450	0.11548	0.11844	0.12343	0.13053	0.13995
1.10	0.11573	0.11676	0.11985	0.12509	0.13256	0.14247
1.11	0.11697	0.11804	0.12128	0.12677	0.13464	0.14515
1.12	0.11821	0.11933	0.12272	0.12848	0.13675	0.14789
1.13	0.11946	0.12062	0.12418	0.13021	0.13891	0.15070
1.14	0.12071	0.12192	0.12565	0.13197	0.14112	0.15358
1.15	0.12197	0.12324	0.12713	0.13375	0.14338	0.15655
1.16	0.12324	0.12457	0.12863	0.13557	0.14568	0.15962
1.17	0.12452	0.12590	0.13015	0.13741	0.14804	0.16278
1.18	0.12581	0.12725	0.13168	0.13928	0.15046	0.16605
1.19	0.12710	0.12861	0.13323	0.14118	0.15294	0.16944
1.20	0.12840	0.12998	0.13480	0.14312	0.15548	0.17296
1.21	0.12971	0.13135	0.13639	0.14509	0.15809	0.1766
1.22	0.13103	0.13274	0.13799	0.14710	0.16076	0.1804
1.23	0.13235	0.13414	0.13961	0.14914	0.16351	0.1843
1.24	0.13368	0.13554	0.14125	0.15122	0.16633	0.1884
1.25	0.13502	0.13696	0.14292	0.15334	0.16924	0.1927
1.26	0.13637	0.13839	0.14461	0.15550	0.17224	0.1972
1.27	0.13773	0.13983	0.14631	0.15772	0.17533	0.2019
1.28	0.13910	0.14128	0.14804	0.15998	0.17852	0.2069
1.29	0.14048	0.14275	0.14979	0.16229	0.18183	0.2121
1.30	0.14187	0.14424	0.15157	0.16466	0.18526	0.2177
1.31	0.14327	0.14574	0.15339	0.16708	0.18882	0.2238
1.32	0.14468	0.14725	0.15523	0.16956	0.19253	0.2302
1.33	0.14610	0.14877	0.15709	0.17211	0.19638	0.2372
1.34	0.14753	0.15031	0.15898	0.17472	0.20040	0.2446
1.35	0.14898	0.15187	0.16091	0.17740	0.20459	0.2526

the plates and that for the incident wave,

$$\phi=\exp ik(x\cos\theta_1+z\cos\theta_3)$$
$$=\exp ik\sin\theta_2(x\sin\beta+z\cos\beta), \quad (40)$$

where β is the angle included between the z-axis and the projection of OP on the xz-plane. The differential equation and the boundary conditions are identical with those described in Sec. 2, except for the fact that in the present case $k\sin\theta_2$ takes the place of k and β takes the place of the angle θ. The formulas of Carlson and Heins may then be used in the general case, provided those substitutions are made.

It is interesting to calculate the magnitude of the reflection coefficient of unconstrained rays for the special case $\epsilon=0$. These rays are incident in the yz-plane; therefore, $\beta=0$. Their angle of incidence is $\psi=\theta_3=90°-\theta_2$. In Eq. (30) we replace θ by 0, k by $k'=k\sin\theta_2=k\cos\psi$, and κ by κ', where κ' is associated with k' in the same way as κ is associated with k; i.e., $k'^2=\kappa'^2+\pi^2/a^2$. It follows that

$$k^2\sin^2\psi=\kappa^2-\kappa'^2. \quad (41)$$

Let ψ' be the angle of refraction associated with the angle of incidence ψ according to Snell's law; i.e.,

$$n=\kappa/k=\sin\psi/\sin\psi'. \quad (42)$$

From Eqs. (41) and (42) then follows $\kappa'=\kappa\cos\psi'$, and

$$\rho=\frac{k'-\kappa'}{k'+\kappa'}=\frac{k\cos\psi-\kappa\cos\psi'}{k\cos\psi+\kappa\cos\psi'}. \quad (43)$$

Repeated use of Eq. (42) leads to

$$\rho=\sin(\psi'-\psi)/\sin(\psi'+\psi). \quad (44)$$

which is identical with one of Fresnel's formulas.

PART II. EXPERIMENTS

The experiments described here are a part of a program undertaken with the purpose of testing the predictions of the theory developed in Part I. In this article we restrict ourselves to the results of experimentation at the Naval Research Laboratory.[12] These include measurements of the magnitude of the reflection coefficient at normal incidence and measurements of the phase of the transmission coefficient with the angle of incidence varying from 0° to 25°.

TABLE II. Tables of $\rho''/2\pi$.

$x=\frac{2a}{\lambda}$	Angle of incidence 0°	5°	10°	15°	20°	25°
1.01	0.51670	0.51693	0.51761	0.51877	0.52044	0.52269
1.02	0.52371	0.52404	0.52503	0.52671	0.52915	0.53244
1.03	0.52916	0.52958	0.53082	0.53295	0.53605	0.54022
1.04	0.53381	0.53430	0.53578	0.53832	0.54200	0.54699
1.05	0.53794	0.53851	0.54021	0.54313	0.54738	0.55315
1.06	0.54173	0.54236	0.54428	0.54758	0.55239	0.55895
1.07	0.54525	0.54596	0.54810	0.55177	0.55713	0.56448
1.08	0.54857	0.54935	0.55171	0.55575	0.56167	0.56981
1.09	0.55173	0.55258	0.55514	0.55957	0.56605	0.57500
1.10	0.55475	0.55567	0.55845	0.56326	0.57032	0.58009
1.11	0.55766	0.55865	0.56165	0.56684	0.57450	0.58512
1.12	0.56047	0.56153	0.56476	0.57034	0.57860	0.59011
1.13	0.56319	0.56433	0.56778	0.57378	0.58266	0.59510
1.14	0.56585	0.56706	0.57075	0.57716	0.58669	0.60010
1.15	0.56844	0.56973	0.57366	0.58050	0.59071	0.60512
1.16	0.57098	0.57235	0.57653	0.58381	0.59471	0.61019
1.17	0.57347	0.57492	0.57936	0.58709	0.59871	0.61531
1.18	0.57592	0.57745	0.58215	0.59035	0.60271	0.62050
1.19	0.57833	0.57994	0.58491	0.59360	0.60675	0.62577
1.20	0.58070	0.58240	0.58765	0.59684	0.61083	0.63115
1.21	0.58304	0.58484	0.59037	0.60009	0.61494	0.63666
1.22	0.58536	0.58726	0.59309	0.60335	0.61909	0.64230
1.23	0.58766	0.58964	0.59579	0.60661	0.62329	0.64809
1.24	0.58993	0.59202	0.59848	0.60989	0.62756	0.65406
1.25	0.59218	0.59438	0.60117	0.61319	0.63191	0.66024
1.26	0.59442	0.59672	0.60384	0.61652	0.63634	0.66665
1.27	0.59664	0.59905	0.60652	0.61988	0.64086	0.67329
1.28	0.59884	0.60137	0.60921	0.62326	0.64547	0.68022
1.29	0.60104	0.60369	0.61190	0.62669	0.65019	0.68749
1.30	0.60323	0.60599	0.61460	0.63016	0.65506	0.69514

[12] The experimental program is described in greater detail in "Physical Optics of Metal Plate Media. II," NRL Report 3599, Washington (1950).

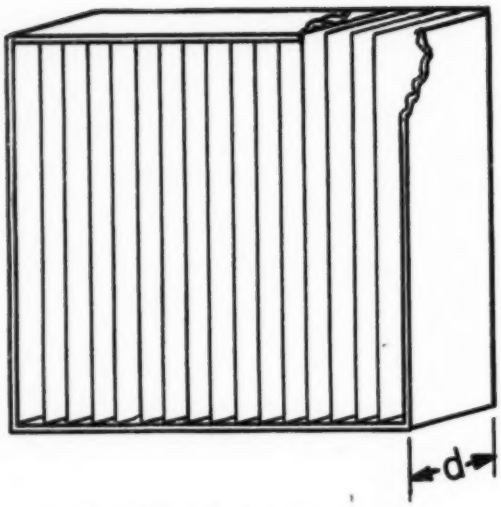


FIG. 6. Sketch of the metal-plate slab.

6. Description of Equipment

The general reflection and transmission measurements were performed on slabs of metal-plate media such as shown in Fig. 6. The large dimensions facing the transmitter were in all cases about 30×30 cm (12″×12″). The plate separation, a, the depth of the slab, d; and the thickness of the plates, t, varied from slab to slab as shown in Table III. A special set of three metal plate slabs was made for the purpose of studying the effect of plate thickness. Their depths and plate separations were 4.00 cm and 2.00 cm, respectively, the plate thicknesses being 0.166 cm, 0.240 cm, and 0.313 cm. The surface of every slab was perpendicular to the plates; i.e., $\epsilon=0$. In spite of all precautions it was not possible to keep the brass plates perfectly straight and for this reason the plate separation varied from point to point, the deviations reaching ±0.015 cm from the design value.

A specially built frequency-stabilized, square-wave-modulated signal generator was the source of radiation. It operated between 2.9 and 3.6 cm, providing a $2a/\lambda$ ratio from 1.03 to 1.28 for slabs 1, 2, and 3 and a ratio from 1.11 to 1.38 for slab 4. A bolometer or crystal served as a detector in conjunction with a Radiation Laboratory Mark IV amplifier and a Ballantine voltmeter.

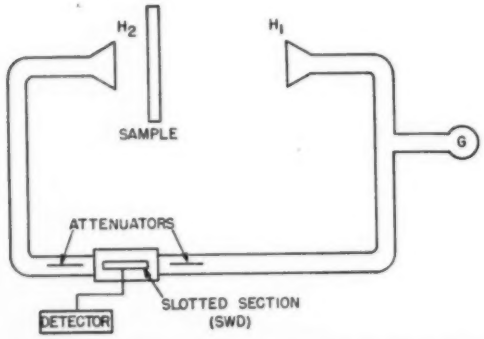


FIG. 7. Apparatus for the measurement of phase change in transmission.

The reflection measurements were made on a modified Michelson interferometer.[13,14] The phase change in transmission through a metal-plate slab was measured on a free-space dielectrometer whose schematic diagram is shown in Fig. 7. A signal is radiated from horn H_1, received by horn H_2, and transmitted through a wave guide containing attenuators and a standing wave detector. A reference signal propagating in the opposite direction in the guide combines with the first signal to produce a standing wave pattern. The attenuators are adjustable so that a high standing wave ratio may be produced in the slotted section. When a slab is interposed between the horns, the phase of the first signal is altered and the minima in the standing wave detector are shifted. From this shift the change of the phase can be calculated.

7. Measurement of the Reflection Coefficient

The primary quantities derived from the theory are the transmission and reflection coefficients of the air-metal-plate interface. These quantities are not directly measurable on the slabs available for the experiment, since reflections from the back surface of the slabs combine with those from the front surface. The resulting multiple reflections lead to a transmission and a reflection coefficient for the entire slab which depend on the depth (thickness) of the slab. For the magnitude of the reflection coefficient the following expression was found:

TABLE III. Dimensions of the metal-plate slabs in centimeters.

Slab No.	d	a	t
1	3.20	1.847	0.080
2	4.80	1.847	0.080
3	9.60	1.847	0.080
4	4.00	2.000	0.075

$$|R| = 2\rho|\sin\Psi|/[(1-\rho^2)^2+4\rho^2\sin^2\Psi]^{\frac{1}{2}}, \quad (45)$$

where $\Psi=\rho''+\delta_2$, and $\delta_2=2\pi nd/\lambda$ is the phase delay between the front and back surfaces of the slab.[9] All quantities entering into the expression for $|R|$ are functions of d or of $2a/\lambda$. Values of $|R|$ calculated for the four experimental slabs of Table III, are presented as the solid curves in Fig. 8.

To emphasize the role of the phase change occurring at the air-metal-plate medium interface, $|R|$ was also calculated assuming, in analogy to true dielectric materials, that no such phase change takes place. This hypothesis is equivalent to setting $\rho''=\pi$. It results in the dotted curves of Fig. 8(a) and (b).

[13] B. A. Lengyel, Proc. Inst. Radio Engrs. **37**, 1242–1244 (1949).

[14] B. A. Lengyel and A. J. Simmons, "An Interferometer for Microwaves," NRL Report 3562 (1949).

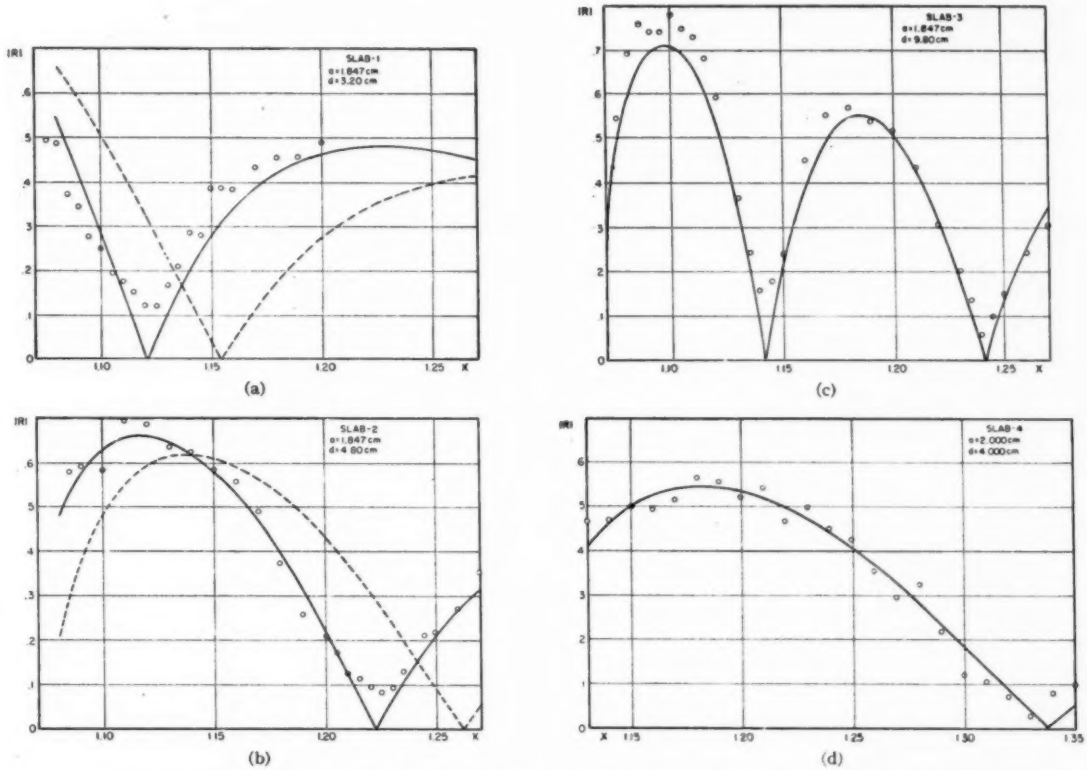


FIG. 8. Reflection from slabs at normal incidence as a function of $x=2a/\lambda$. (Solid line: $|R|$ calculated from the Carlson-Heins theory. Dotted line: $|R|$ calculated neglecting phase change on interface.)

Experimental values for $|R|$ were obtained for each of the four slabs. The quantity $x=2a/\lambda$ was varied by changing λ, and $|R|$ was measured as a function of x. The results, plotted as small circles in Fig. 8, are in obvious disagreement with the dotted curves, but in reasonable agreement with the solid curves calculated from the Carlson-Heins theory.

In appraising the agreement of experimental values of $|R|$ with those deduced from the theory, one must keep in mind the deviations of the metal-plate structure from the mathematical idealization and the errors inherent in the technique of measurements. The finite conductivity of the plates is a source of loss within the slab. When this loss is taken into account, the solid curves in Fig. 8 will no longer touch the axis. This effect, however, is negligible, since the loss is only about 0.06 db/meter. The principal systematic errors inherent in the measurement technique are those caused by diffraction, by the divergent character of the radiation, and by the presence of unwanted reflections. By careful choice of the dimensions of the apparatus the effect of diffraction was reduced to negligible proportions. The divergence of the radiation introduced a spread in the angle of incidence amounting to $\pm 6°$ in the extreme. The effect due to this spread is very small, since the reflection coefficient is not sensitive to angle near normal incidence. The largest error is caused by reflection from the carriage of the interferometer. The error in $|R|$ due to a carriage reflection of magnitude α is approximately $\alpha|1+R\exp i\phi|$, where ϕ depends on the position of the slab with respect to that of the normalizing solid plate.[14] Since the phase of R varies slowly with λ, this error is a slowly varying function of the wavelength. One may expect that in certain wavelength regions the measured value of $|R|$ will lie consistently above, in other regions below its correct value. The carriage reflections of the interferometer used remained below 0.05 over the entire 3-cm band; therefore, the average error of the measurement of $|R|$ was less than 0.05 and the maximum error less than 0.10.

8. The Measurement of Phase Change in Transmission

We measured the change of phase of a signal propagating between two horns (Fig. 7) when a slab of metal-plate medium was interposed. The theoretical expression for this quantity is[9]

$$T'=\rho'+\Psi'-\pi-\delta_1, \tag{46}$$

where $\delta_1=2\pi d\cos\psi/\lambda$ and

$$\tan\Psi'=[(1+\rho^2)/1-\rho^2]\tan\Psi. \tag{47}$$

The experimental results obtained at normal incidence are presented in Fig. 9. Samples of the data obtained at oblique incidence in the plane of maximum constraint

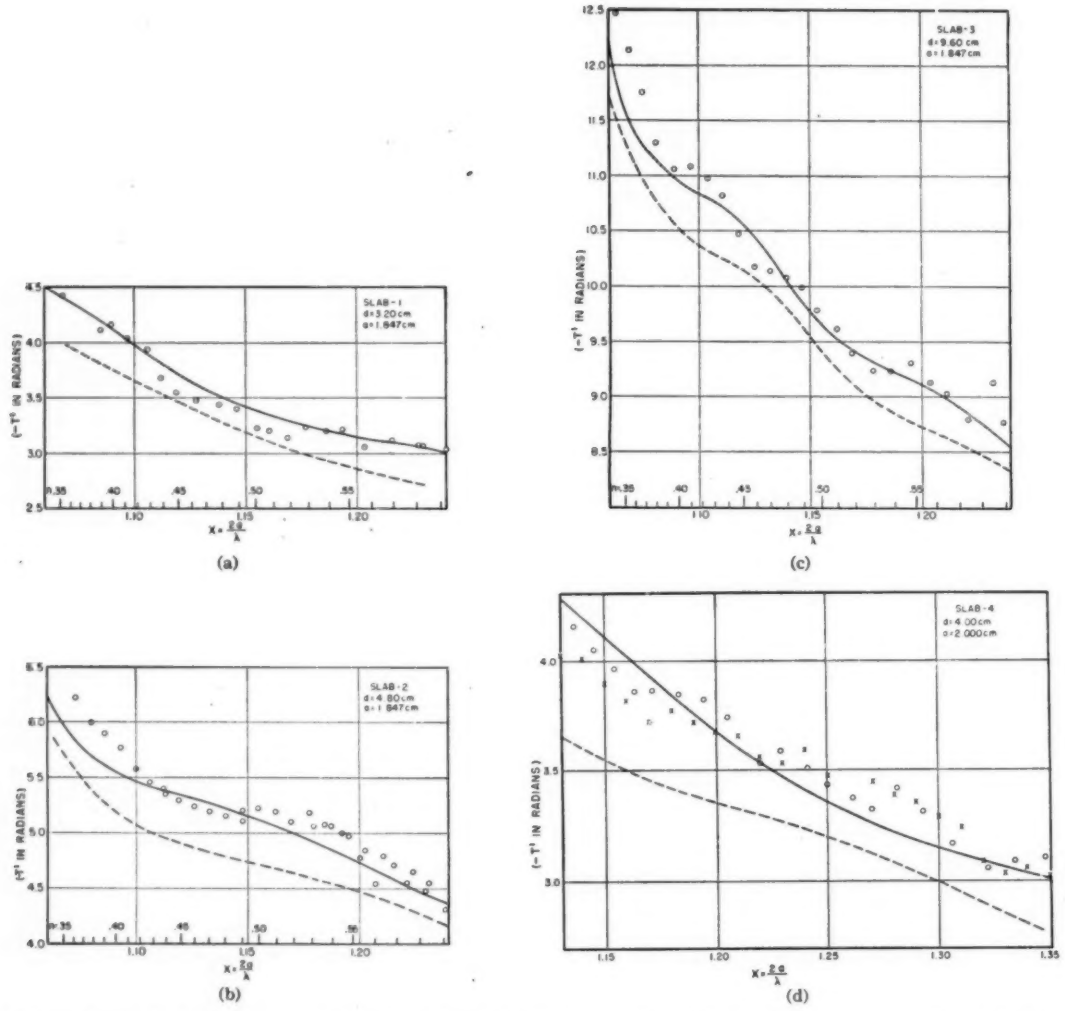


FIG. 9. Phase change in transmission at normal incidence ($\epsilon=0$). o, experimental, thick spacers. x, experimental, thin spacers. (Solid line: T' calculated from the Carlson-Heins theory. Dotted line: T' calculated neglecting phase change on interface.)

are shown in Fig. 10. The values of T' calculated from Eq. (46) are the solid curves. The dotted curves in Fig. 9 represent T' calculated assuming $\rho''=\pi$, and again the experimental data demonstrate that this hypothesis is not tenable.

In this type of measurement the principal sources of error are the mechanical imperfections of the slab and the presence of unwanted reflections. Transmitting and receiving horns were only about 60 cm apart; therefore, diffraction effects were negligible. The divergence of the beam on the other hand was greater than in the case of the reflection measurements. Multiple reflections between horns and the slab and internal reflections within the closed system (consisting of wave guide sections, attenuators, and horns) were the most disturbing factors and a greater part of our effort was directed toward their reduction. When the measurements were made at normal incidence, each reading was repeated with the slab moved a quarter-wavelength, and the average phase change was recorded. At oblique incidence readings were taken on both sides of the normal; their averages appear on Fig. 10. The internal consistency of the experimental data might be appraised by examining Fig. 9d. The experimental points marked by circles were obtained first, then the spacers on the slab were exchanged for new and thinner ones, an attenuator in the dielectrometer was exchanged for one better matched, and the measurements were repeated. The experimental points finally obtained are marked by crosses. The mean deviation from the theoretical values is 0.10 radian or about 3 percent of the measured phase change.

While no mathematical theory of the effect of plate thickness exists at the present, physical intuition leads one to expect that formula (30) for the magnitude of the reflection coefficient remains approximately valid for moderate values of the plate thickness, t, provided that the plate separation a, is used in the computation of κ. This conjecture is supported by the data of Fig. 11, where the theoretical reflection coefficient for normal incidence is plotted for a slab with $a=2.00$ cm and

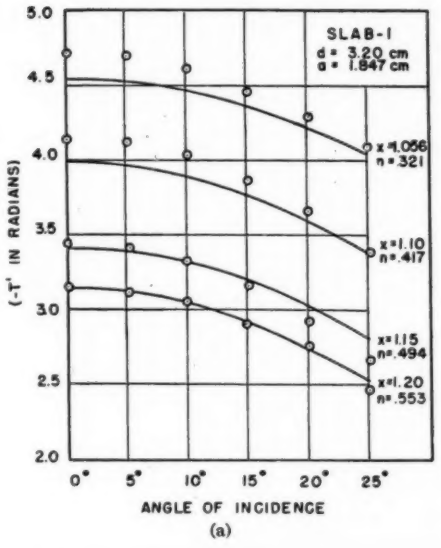


FIG. 10. Phase change in transmission at oblique incidence ($\epsilon=0$).

$d=4.00$ cm, together with experimental points obtained for $t=0.166$, 0.240, and 0.313 cm.

9. Summary of Experimental Results

At normal incidence both amplitude and phase predictions of the Carlson-Heins theory have been verified experimentally for the range $n=0.35$ to $n=0.67$. Phase measurements were made over the same range of n with the angle of incidence ranging from 0 to 25°; the plane of incidence being the plane of maximum constraint. The results agreed essentially with the predictions of the theory.

10. Acknowledgments

The measurements were made by Messrs. A. J. Simmons and J. K. Burton of the Naval Research Laboratory. Mr. Simmons designed most of the experimental equipment. Dr. J. E. Eaton contributed constructive suggestions pertaining to the mathematical sections of this paper.

APPENDIX A. THE ROOTS OF $\cos p-\cos q=0$

According to the defining Eqs. (9) and (10), $p(w)=a(k^2-w^2)^{\frac{1}{2}}$, $q(w)=ak\sec\epsilon\sin\psi-wa\tan\epsilon$. While the sign of p has no real significance in the equation to be solved, the function can be made single-valued by slitting the w plane from $-\infty$ to $-k$ and from k to ∞ along (or close to) the real axis and letting $p(0)=ak$. For a real k this definition implies $p(\bar{w})=\bar{p}(w)$, the bar indicating the complex conjugate.

The transcendental equation

$$\cos p-\cos q=0, \tag{A-1}$$

may be replaced by the following two sets of algebraic equations:

$$p-q=2\pi m, \tag{A-2}$$

$$p+q=2\pi m, \tag{A-3}$$

where m runs through all integers.

When k is real, and this is the case which is of ultimate interest, the real solutions of Eqs. (A-2) and (A-3) can be found geometrically. They are the abscissas of the intersections of the circle $p/a=(k^2-w^2)^{\frac{1}{2}}$ with the set of parallel lines, one of which is $q/a=k\sec\epsilon\sin\psi-w\tan\epsilon$, the distance of adjacent lines being $2\pi\cos\epsilon/a$. According to Eq. (5) this distance is greater than k, the radius of the circle; therefore, at most two lines will intersect the circle. Figure 12 shows a typical situation.

In order to apply the technique of Fourier transform it is necessary to permit k to be complex with a small positive imaginary part (Appendix B). To solve Eq. (A-1) in this general case, we write $w=k\cos(\psi'-\epsilon)$ and determine ψ'. Then, $p(w)=\pm ak\sin(\psi'-\epsilon)$, and

$$\begin{aligned} q(w)&=ak\sec\epsilon[\sin\psi-\sin\epsilon\cos(\psi'-\epsilon)]\\ &=ak\sec\epsilon[\sin\psi-\sin\psi'+\sin(\psi'-\epsilon)\cos\epsilon]\\ &=ak\sec\epsilon(\sin\psi-\sin\psi')\pm p(w). \end{aligned}$$

Whenever ψ' satisfies Eq. (18) then $w=k\cos(\psi'-\epsilon)$ satisfies Eq. (A-1). When ψ' satisfies Eq. (18), $180°-\psi'$ also satisfies it, and another solution is $k\cos(180°-\psi'-\epsilon)=-k\cos(\psi'+\epsilon)$. For $m=0$ the solutions[15] are $w_1=k\cos(\psi-\epsilon)$ and $w_2=-k\cos(\psi+\epsilon)$. This pair is real for real k and will therefore be called essentially real. Other essentially real solutions may exist for $m=\pm1$. If there are any, we select ψ' so that for $\mathrm{Im}\,k=0$, $\cos\psi'<0$, then $\cos(180°-\psi')>0$; finally we define $w_3=k\cos(180°-\psi'-\epsilon)=-k\cos(\psi'+\epsilon)$ and $w_4=k\cos(\psi'-\epsilon)$.

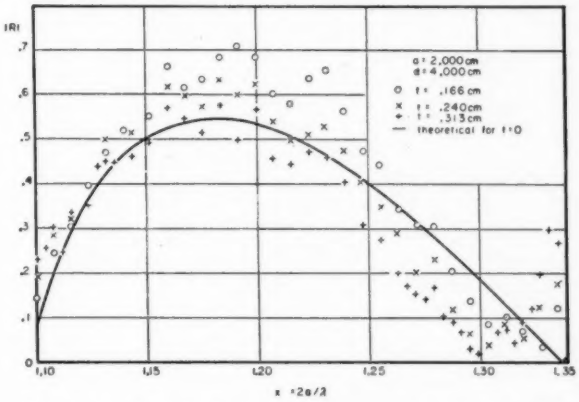


FIG. 11. The effect of plate thickness on the magnitude of the reflection coefficient.

[15] In terms of ψ' there are additional solutions of Eq. (A-1), namely, $\psi=2\epsilon-\psi$ and $\psi'=2\epsilon+\psi-180°$; but these lead to the same values of w.

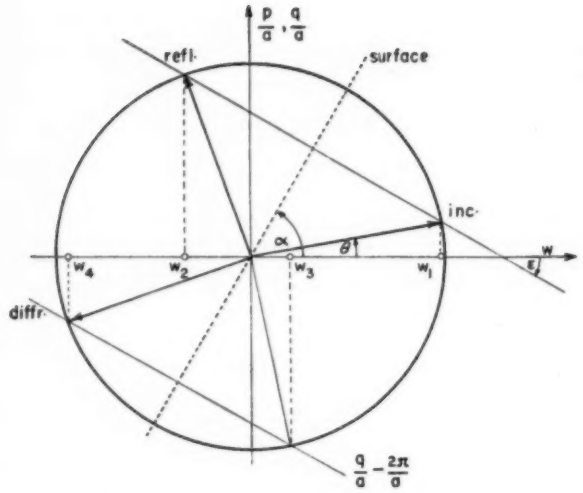


FIG. 12. Geometrical solution of the equation $\cos p = \cos q$.

It is finally necessary to prove that the complex (not essentially real) roots of Eq. (A-1) are simple and occur in conjugate pairs. The passage from Eq. (A-1) to Eqs. (A-2) and (A-3) does not alter the multiplicities and the roots of the latter systems have the required properties. In fact, a root of Eq. (A-2) of multiplicity higher than one must satisfy

$$(dp/dw) - dq/dw = -[aw/(k^2-w^2)^{\frac{1}{2}}] - a\tan\epsilon = 0; \quad \text{(A-4)}$$

therefore, $w = -k\sin\epsilon$, which is real for real k. For the same reason the complex roots of Eq. (A-3) are also simple. Let w be a root of Eq. (A-2); then $p(\bar{w}) - q(\bar{w}) = \bar{p}(w) - \bar{q}(w) = 2\pi m$; $\bar{w}$ is also a root. The same applies to Eq. (A-3).

APPENDIX B. THE CURVE C AND RELATED MATTERS

The expression (14) derived for the total field by Carlson and Heins must be justified for the case in which the equation $\cos p = \cos q$ has essentially real roots other than w_1 and w_2 and the location of the path C with respect to these additional roots must be determined. The procedure of Carlson and Heins requires that k be assumed complex, $k = k_r + ik_i$, where k_i is positive and is ultimately allowed to tend to zero. It is then shown that the Fourier transforms of all terms in the integral equation to be solved (reference 1, p. 319) have a common strip of regularity; it is, therefore, permissible to apply the Fourier transform technique. The transform of the kernel of the integral equation is $f(w)$ its singularities depend on the solutions of $\cos p = \cos q$. When w_1 and w_2 are the only essentially real poles of $f(w)$, Carlson and Heins have shown that the common domain of regularity of the transforms is $\mathrm{Im}w_2 < \mathrm{Im}w < \mathrm{Im}w_1$ and that C is any curve in this strip extending to infinity on both sides. We will show that in the general case at least one of the additional, essentially real roots, namely, w_3, lies in this strip and that $\mathrm{Im}w_3 > \mathrm{Im}w_4$. It is therefore possible to specify a new strip of regularity as follows: $\mathrm{Im}w_3 > \mathrm{Im}w > \max(\mathrm{Im}w_2, \mathrm{Im}w_4)$. The curve C must be led through this strip leaving w_1 and w_3 in the upper, w_2 and w_4 in the lower half-plane. It may then be deformed in the usual manner and k_i may be allowed to approach zero. All conclusions of Carlson and Heins may then be justified using this new strip of regularity.

In fact, from Eq. (18) follows

$$\mathrm{Re}\sin\psi' = \sin\psi + rk_r/|k|^2 \quad \text{(B-1)}$$

$$\mathrm{Im}\sin\psi' = -rk_i/|k|^2, \quad \text{(B-2)}$$

where r is the real number $2\pi m/a$ secϵ. It follows from the assumption about the magnitude of r [see Eq. (5)] that $r\,\mathrm{Re}\sin\psi' > 0$, and from Eq. (B-2) that $r\,\mathrm{Im}\sin\psi' < 0$. From these relations we may conclude that for small positive values of k_i the real and imaginary parts of $\cos\psi'$ have the same sign. According to its definition $\cos\psi'$ is negative for $k_i = 0$; therefore, by continuity its real and imaginary parts are negative for small positive values of k_i.

$$\begin{aligned}\mathrm{Im}w_3 - \mathrm{Im}w_4 &= -\mathrm{Im}[k\cos(\psi'+\epsilon) + k\cos(\psi'-\epsilon)] \\ &= -2\,\mathrm{Im}k\cos\psi'\cos\epsilon \\ &= -2(k_i\,\mathrm{Re}\cos\psi' + k_r\,\mathrm{Im}\cos\psi')\cos\epsilon. \quad \text{(B-3)}\end{aligned}$$

This is positive for a small positive k_i. To complete the proof of our propositions it is now sufficient to verify that

$$\mathrm{Im}w_1 > \mathrm{Im}\tfrac{1}{2}(w_3 + w_4) > \mathrm{Im}w_2. \quad \text{(B-4)}$$

Let A, B, and C denote the three terms in Eq. (B-4); then

$$A = k_i\cos(\psi - \epsilon) = k_i\cos\psi\cos\epsilon + k_i\sin\psi\sin\epsilon, \quad \text{(B-5)}$$

$$\begin{aligned}B &= \tfrac{1}{2}\,\mathrm{Im}[k\cos(\psi'-\epsilon) - k\cos(\psi'+\epsilon)] \\ &= \mathrm{Im}k\sin\psi'\sin\epsilon \\ &= (k_r\,\mathrm{Im}\sin\psi' + k_i\,\mathrm{Re}\sin\psi')\sin\epsilon \\ &= k_i\sin\psi\sin\epsilon \quad \text{(B-6)}\end{aligned}$$

$$C = -k_i\cos(\psi+\epsilon) = -k_i\cos\psi\cos\epsilon + k_i\sin\psi\sin\epsilon. \quad \text{(B-7)}$$

Therefore, $A - B = B - C = k_i\cos\psi\cos\epsilon > 0$, which establishes (B-4).

Experimental Determination of the Reflection Coefficient of Metal-Plate Media*

J. Ruze and M. Young
Air Force Cambridge Research Laboratories, Cambridge, Massachusetts
(Received August 9, 1950)

A method of measuring the reflection coefficient of metal-plate media at 3 cm is described. The technique differs from that presented in the preceding paper in that the measurements are made in a closed system and back reflections are eliminated by an absorbing wedge. Angle of incidence and the plate thickness are varied. The agreement with the Carlson-Heins theory is excellent for very thin plates. For plates of moderate thickness near normal incidence the measured reflection is higher than that predicted for an infinitely thin set of plates with identical index of refraction.

THE reflection coefficient of a stack of parallel metal plates has been determined by comparison of the relative energy reflected from the test body with that from a perfectly conducting sheet. In these experiments the electromagnetic field was confined to a pair of parallel semicircular sheets. A transmitting and a receiving horn, aimed at the center of the circle, were rotatable about the circular arc of the plates. The metal-plate interface, fastened along the diameters of the plates, was backed by a tapered wooden plug. This plug acted as an absorber of the energy transmitted through the interface. Figures 1 and 2 show the equipment and some of the test sections used.

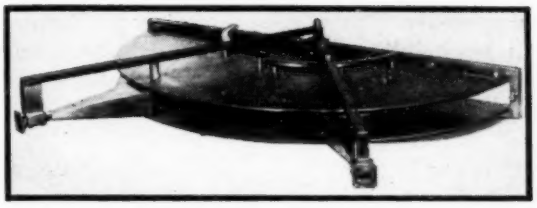


FIG. 1. Reflectometer.

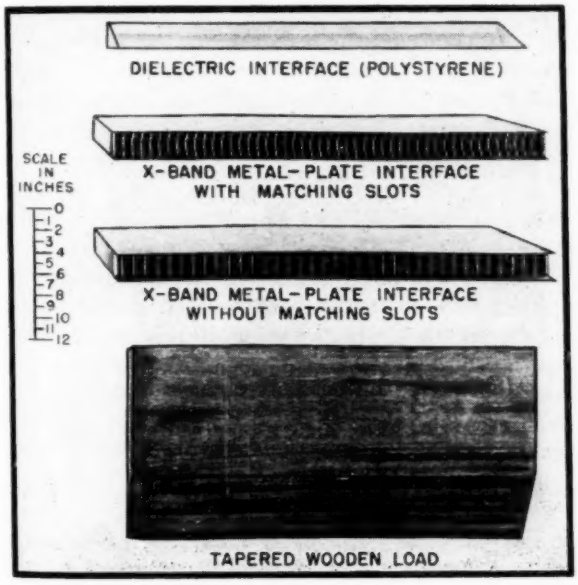


FIG. 2. Lens test sections.

To check the accuracy of the device, data were first obtained on a polystyrene interface at a wavelength of 3.3 cm (Fig. 3). It should be observed that the data agree comparatively well with the theoretical fresnel equations. The angle of incidence is limited to about 70° because of direct radiation between the horns.

Measurements were then obtained from metal-plate media at various plate spacings and different plate thicknesses. Typical experimental data together with the theoretical curve computed by Lengyel[1] are indicated in Fig. 4. It is to be noted that the predicted diffraction beam is well reproduced; however, the reflection at around normal incidence is higher than predicted. This result was obtained from all measurements made with the exception of those on extremely thin

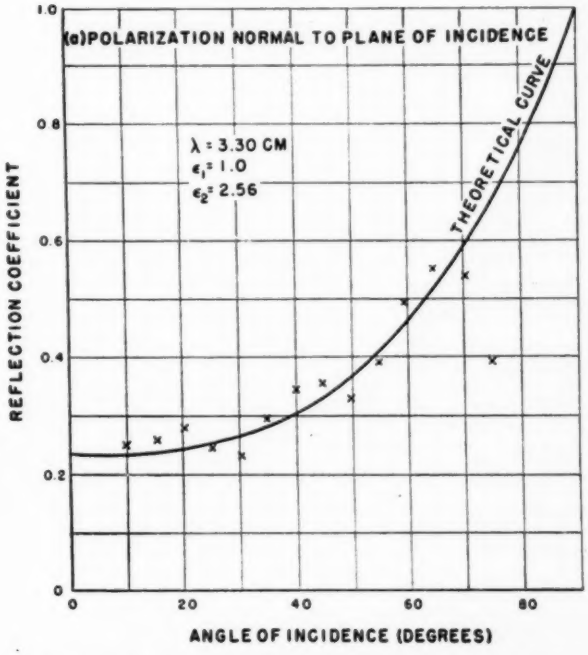


FIG. 3. Reflection coefficient of air-polystyrene interface.

* Based on Cambridge Field Station, Air Materiel Command Report No. E4056, of the same title, dated April, 1949.

[1] B. A. Lengyel, J. App. Phys. 22, 265 (1951).

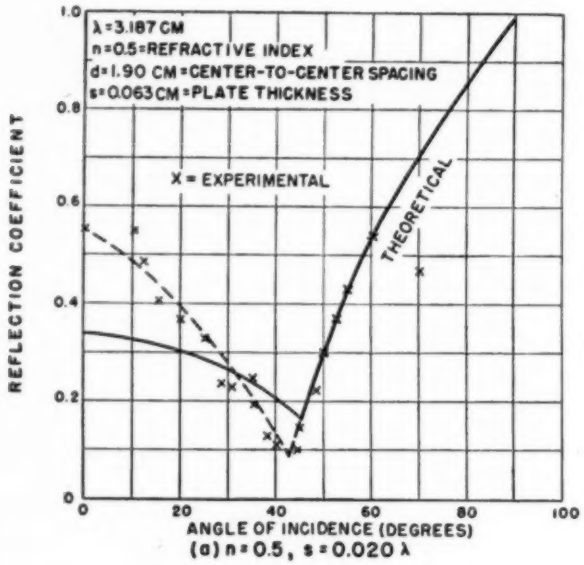


FIG. 4. Reflection from metal plates, polarization in plane of plates.

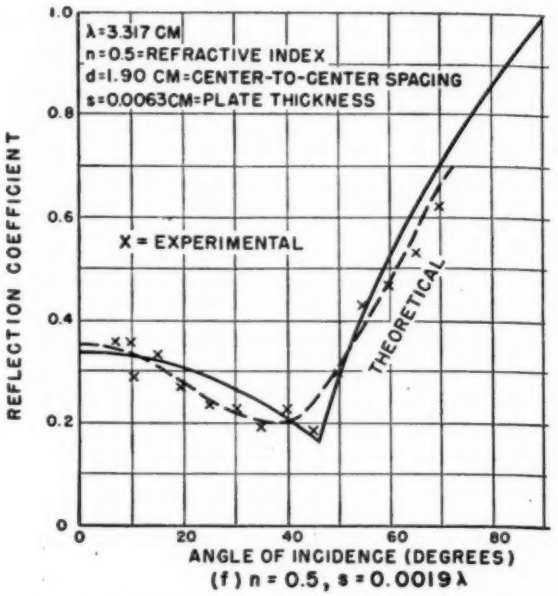


FIG. 5. Reflection from metal plates, polarization in plane of plates.

plates formed by supporting metal foil in blocks of Polyfoam, the experimental data of which are shown in Fig. 5.

An attempt was made to determine the magnitude and phase of the reflection coefficient by exploring the field standing wave distribution in the immediate proximity of the metal-plate media. No reliable data could be obtained in this attempt because of difficulty with poor contact between the pick-up probe and the metal sheets.

This work was performed in the Antenna Laboratory of the Air Force Cambridge Research Laboratories at Cambridge, Massachusetts.

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

On Amplitude Bounds for Certain Relaxation Oscillations

NELSON WAX
University of Illinois, Urbana, Illinois
(Received July 28, 1950)

Amplitude bounds are obtained for the unique periodic solution of the generalized Lienard equation $\ddot{x}+f(x)\dot{x}+g(x)=0$, when the functions $f(x)$ and $g(x)$ are restricted suitably. The restrictions are less severe for the lower than for the upper bounds. A class of equations is exhibited which includes the van der Pol equation and satisfies all the restrictions. Numerical results are given for the van der Pol equation.

I. INTRODUCTION

THE amplitude of oscillation is of considerable importance in the application of the theory of self-excited relaxation oscillators to physical problems. The existence and uniqueness of a periodic solution for a large class of equations of relaxation oscillations has been established by a number of workers,[1,2,3d] but the proofs have been nonconstructive existence proofs. The amplitude usually has been found after the detailed behavior of the periodic solution was obtained by graphical, numerical, or analytic methods.[3] Minorsky[4] has emphasized the difficulties of convergence which are encountered in trying to obtain the periodic solution of the van der Pol equation*

$$\ddot{x}+\mu(x^2-1)\dot{x}+x=0, \tag{1.1}$$

[1] N. Levinson and O. K. Smith, Duke Math. J. 9, 382 (1942).
[2] S. Lefschetz, *Lectures on Differential Equations* (Princeton University Press, Princeton, New Jersey, 1946).
[3] See (a) N. Minorsky, *Introduction to Non-Linear Mechanics* (Edward Brothers, Inc., Ann Arbor, Mich., 1947); (b) A. A. Andronow and C. E. Chaikin, *Theory of Oscillations* (Princeton University Press, Princeton, New Jersey, 1949) English language edition; (c) N. Kryloff and N. Bogoliuboff, *Introduction to Non-Linear Mechanics* (Princeton University Press, Princeton, New Jersey, 1947) free translation by S. Lefschetz; (d) J. J. Stoker, *Nonlinear Vibrations in Mechanical and Electrical Systems* (Interscience Publishers, Inc., New York, 1950). References (a) and (d) contain bibliographies.
[4] N. Minorsky, J. Franklin Inst. 248, 205 (1949).
* A dot indicates differentiation with respect to t (the time), a dash, differentiation with respect to x.

(μ const.>0) analytically, for arbitrary μ. Similar difficulties arise for other equations of relaxation oscillations.

It seems desirable to have estimates of the amplitude in the absence of precise information concerning the behavior of the periodic solution. LaSalle[5] has studied a class of differential equations, which includes Eq. (1.1), and has obtained bounds for both the amplitude and period of the periodic solutions. A somewhat different, but overlapping, class of equations is treated here.

It is the purpose of this note to show that amplitude bounds may be obtained quite simply for the periodic solution of the generalized Lienard equation

$$\ddot{x}+f(x)\dot{x}+g(x)=0, \tag{1.2}$$

when the functions $f(x)$ and $g(x)$ are restricted suitably. All of the restrictions on these functions are listed in the next paragraph.

It is assumed throughout that $g(x)$ is an odd differentiable function of x, with $g(x)>0$ if $x>0$, $f(x)$ is an even continuous function for which there exists an $x_0>0$ such that $f(x)<0$ when $|x|<x_0$, $f(x)>0$ for $|x|>x_0$. Let the odd function, $F(x)$, be defined by

$$F(x)=\int_0^x f(\xi)d\xi,$$

and the even function, $G(x)$, by

$$G(x)=\int_0^x g(\xi)d\xi.$$

It is assumed that $F(\infty)=G(\infty)=\infty$. There will exist then, an $\alpha>0$ such that $F(x)<0$ if $0<x<\alpha$, and $F(x)>0$ and strictly increasing when $x>\alpha$. Lower bounds are obtained here if $F^2(x_0)>2G(\alpha)$, and upper bounds if $-[g(x)/f(x)]$ is monotone increasing when $x<-x_0$. The $f(x)$ is less, and the $g(x)$ more general than the corresponding functions considered by LaSalle.

A summary of previous work occupies the remainder of this Section. These remarks are developed further in Sec. II, and the role of the function $-[g(x)/f(x)]$ is investigated. Lower bounds for the amplitude are given in Sec. III, upper bounds in Sec. IV. Section V contains some examples and numerical results for the van der Pol equation.

Considerable use will be made of results first given by Lienard; this work has been generalized by Levinson and Smith, and it is a portion of their treatment which is summarized immediately below.

Equation (1.2) can be written in the form

$$dv/dx=-f(x)-[g(x)/v], \tag{1.3}$$

where $v=\dot{x}$, and the solutions of Eq. (1.2) studied in the phase plane. A curve in the phase plane, corresponding to a solution of Eq. (1.2), will be termed an integral curve, a phase trajectory, or an orbit interchangeably.†

[5] J. LaSalle, Quart. Appl. Math. **7**, 1 (1949). A preliminary account of this work is given in 3(b), Appendix C.

† Reference 1 gives some properties of integral curves.

Equation (1.3) remains unchanged when v and x are replaced by $-v$ and $-x$; closed integral curves, therefore, have mirror symmetry about the origin, in the phase plane. One defines the function $\lambda(x, v)$ [which corresponds to the instantaneous energy of a physical system described by Eq. (2)] by

$$\lambda(x, v)=(v^2/2)+G(x). \tag{1.4}$$

Using Eqs. (1.3) and (1.4), one obtains

$$d\lambda/dx=-f(x)v. \tag{1.5}$$

The contours $\lambda(x, v)=\text{constant}>0$ are closed. Integral curves in the phase plane will cut inward across the contours $\lambda(x, v)=\text{constant}>0$, when $|x|>x_0$, and outward when $|x|<x_0$, since $f(x)>0$ when $|x|>x_0$, and $f(x)<0$ when $|x|<x_0$. All integral curves in the interior of the contour $\lambda(x, v)=G(x_0)$ cut outward across the boundary. All orbits exterior to the closed region bounded by the contour $\lambda(x, v)=G(x_0)$, at time $t=t_1$, remain exterior to it for all $t>t_1$.

If the Lienard coordinates

$$y=v+F(x) \tag{1.6}$$

and x are introduced, then Eq. (1.3) becomes

$$dy/dx=-g(x)/[y-F(x)]. \tag{1.7}$$

One may define the function $\phi(x, y)$ by

$$\phi(x, y)=(y^2/2)+G(x) \tag{1.8}$$

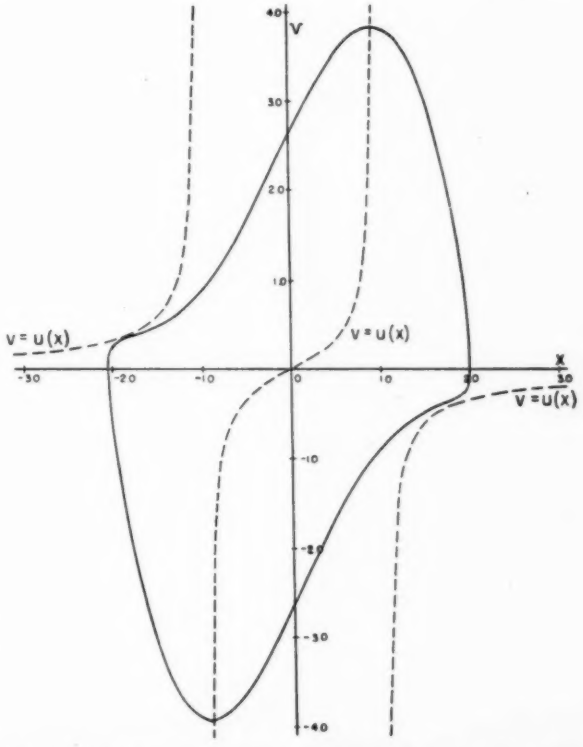


FIG. 1. The periodic solution of Eq. (1.1) with $\mu=100/49$ is shown here in the phase plane. The branches of the function $u(x)$ are also presented.

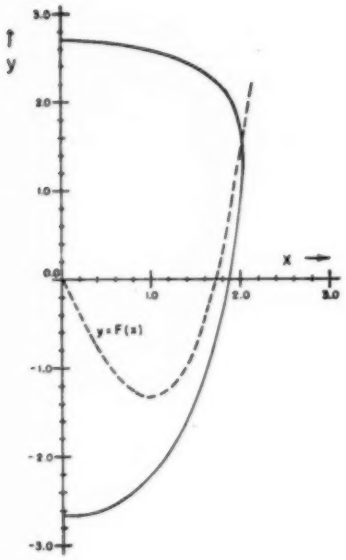


FIG. 2. The right half of the periodic orbit shown in Fig. 1 is given here in the Lienard plane with the graph of $y=F(x)$.

and obtain

$$d\phi/dx=F(x)y'. \tag{1.9}$$

Closed integral curves in the Lienard plane have mirror symmetry about the origin. A unique closed integral curve exists in the Lienard plane for Eq. (1.7); and, therefore, a unique closed integral curve exists in the phase plane for Eq. (1.3). This establishes the existence and uniqueness of the periodic solution of Eq. (1.2) (except for translations in t). The periodic orbit cuts the x axis once and only once in the right half of the phase plane.

Figures 1 and 2 are graphical presentations of a periodic solution of Eq. (1.1) in the phase and Lienard planes, respectively.

II. AUXILIARY REMARKS

It is sufficient to consider the behavior of the periodic orbit in the upper half of the phase plane or the right half of the Lienard plane because of symmetry, and this will be done here unless otherwise stated.

Use is made of the properties of the function $u(x)=-[g(x)/f(x)]$ in establishing bounds. One sees from Eq. (1.3) that $v=u(x)$ represents the contour of zero slope for all integral curves in the phase plane. $u(x)$ is an odd continuous function of x, except for poles at $|x|=x_0$, positive in the intervals $x<-x_0$ and $0<x<x_0$, negative when $-x_0<x<0$ and $x>x_0$. Since dv/dx is a continuous function of x and v, except along the x axis, then integral curves in the upper half of the phase plane will have positive slope when below $v=u(x)$ and negative slope when above $v=u(x)$, for $x<-x_0$. In $-x_0\leq x\leq 0$ integral curves have positive slope in the upper half plane. In $0\leq x\leq x_0$ integral curves have positive slope when above $u(x)$ and negative slope when below $u(x)$; integral curves have negative slope for $x\geq x_0 (v>0)$.

If $u(x)$ is a monotone increasing function of x in the interval $-B\leq x<-x_0$, then $v=u(x)$ is an upper bound of $v(x)$ for all orbits, starting‡ at $(\xi, 0)$, $-B\leq\xi<-x_0$, in this interval.§ Proof: Consider the orbit which starts at $(\xi, 0)$, $-B\leq\xi<-x_0$. It is below the contour $v=u(x)$ for some $x<-x_0$, $v\neq 0$, and, thus, has positive slope there. dv/dx is continuous everywhere except along the x axis, and is positive at $x=-x_0$. Since $dv/dx=0$ on $v=u(x)$, phase trajectories move horizontally to the right away from $v=u(x)$; they have negative slope when above $v=u(x)$. Therefore, if this orbit crosses $v=u(x)$, it must do so at least twice. Let $x=\gamma_1$ be the abscissa of a crossing point, and $x=\gamma_2$ that of the next succeeding crossing point, with $(dv/dx)<0$ in $\gamma_1<x<\gamma_2<-x_0$. It follows that $v(\gamma_1)=u(\gamma_1)>v(\gamma_2)=u(\gamma_2)$, which is contrary to assumption. Therefore, $v(x)\leq u(x)$ for this orbit, in the interval $-B\leq x<-x_0$, and, hence, for all orbits starting on the x axis in the same interval.

The amplitude, A, of the periodic orbit is defined by

$$A=\max|x_p|.$$

The periodic orbit will be designated by the subscript "p" hereafter.

Some properties of $\lambda(x, v)$ will be used later. $\lambda(x, v)$ is an even function of both x and v. Let $x=\xi$ be an abscissa common to both of the contours $\lambda(x, v)=k_1$, $\lambda(x, v)=k_2$ where $k_1>k_2>0$. Then, $[v_1^2(\xi)/2]+G(\xi)=k_1>[v_2^2(\xi)/2]+G(\xi)=k_2$, or $|v_1(\xi)|>|v_2(\xi)|$. Therefore, the contour $\lambda(x, v)=k_2$ is interior to the contour $\lambda(x, v)=k_1$. Furthermore, given any finite region in the phase plane, a constant $k>0$ can be found such that the closed contour $\lambda(x, v)=k$ contains the region in its interior.

In the upper half-plane $\lambda_p(x, v)$ is a minimum at $x=-x_0$ and a maximum at $x=x_0$, by Eq. (1.5). $\lambda_p(x, v)$ is strictly decreasing when $x<-x_0$, increases monotonically in the closed interval $[-x_0, x_0]$, and is strictly decreasing when $x>x_0$, with increasing x. $\lambda_p(-A, 0)=\lambda_p(A, 0)$ and, therefore, an upper bound for $v_p(x)$ in the closed interval $[-A, 0]$ is $v_p(x)$ in the closed interval $[0, A]$. Furthermore, the contour $\lambda(x, v)=G(A)$ is an upper bound for $v_p(x)$ in the closed interval $[-A, -x_0]$ and a lower bound in the closed interval $[x_0, A]$.

III. LOWER BOUNDS FOR THE AMPLITUDE

A lower bound for A is obtained implicitly in Levinson and Smith's proof of the existence and uniqueness of the periodic orbit. This bound is $A>\alpha$, where α is the non-zero root of $F(x)=0$. The bound $A>\alpha$ may be improved at times.

If one integrates Eq. (1.3) for the periodic orbit in the second quadrant of the phase plane, between the

‡ Since $\dot{x}=v$, when $v>0$, x increases as t increases, when $v<0$, x decreases as t increases. Integral curves, therefore, encircle the origin clockwise.

§ Reference 5 contains a similar statement.

limits $x=-x_0$ and $x=0$, one obtains

$$v_p(0)-v_p(-x_0)=F(-x_0)-\int_{-x_0}^{0}\frac{g(\xi)d\xi}{v_p(\xi)}. \quad (3.1)$$

Now $v_p(-x_0)>0$, and $v_p(x)$ is positive and bounded. Thus, $|F(x_0)|<v_p(0)$. Now from Eq. (1.9), $\phi_p(0)<\phi_p(x)$ for $0<x\leq\alpha$, in the first quadrant of the Lienard plane. One has then that

$$F^2(x_0)<v_p^2(0)<v_p^2(\alpha)+2G(\alpha). \quad (3.2)$$

An upper bound for $v_p(\alpha)$ is calculated now. A lower bound for $\lambda_p(x, v)$ in $x_0\leq x\leq A$ is $G(A)$ and, therefore,

$$v_p(x)>v_L(x)=2^{\frac{1}{2}}[G(A)-G(x)]^{\frac{1}{2}} \quad (3.3)$$

in $x_0\leq x<A$. An upper bound is obtained by substituting $v_L(x)$ into Eq. (1.3) and integrating. This yields at $x=\alpha$ in the first quadrant of the phase plane

$$v_p(\alpha)<F(A)+2^{\frac{1}{2}}[G(A)-G(\alpha)]^{\frac{1}{2}}. \quad (3.4)$$

The use of inequalities (3.2) and (3.4) gives

$$F^2(x_0)<F^2(A)+2^{\frac{1}{2}}F(A)[G(A)-G(\alpha)]^{\frac{1}{2}}+2G(A). \quad (3.5)$$

A lower bound for A, then, will be a positive real root, $A_{L1}>\alpha$, of

$$F^2(x_0)=F^2(x)+2^{\frac{1}{2}}F(x)[G(x)-G(\alpha)]^{\frac{1}{2}}+2G(x), \quad (3.6)$$

if such a root exists. The right-hand side of Eq. (3.6) is a continuous real positive strictly increasing function of x when $x>\alpha$. Thus, just one root, $A_{L1}>\alpha$ of Eq. (3.6) will exist whenever $F^2(x_0)>2G(\alpha)$.

Let an upper bound for $v_p(0)$ be $v_{Uk}(0)$, (v_{uk} is obtained in the next section), then

$$v_p(x)<[v_{Uk}^2(0)-2G(x)]^{\frac{1}{2}}, \quad (3.7)$$

when $-x_0\leq x<0$. If this upper bound for $v_p(x)$ is substituted into the integral appearing in Eq. (3.1), one obtains

$$v_p(0)>F(-x_0)+v_{Uk}(0)-[v_{Uk}^2(0)-2G(x_0)]^{\frac{1}{2}}=\Delta_k(0). \quad (3.8)$$

The real root $A_{Lk}>\alpha$, of

$$\Delta_k^2(0)=F^2(x)+2^{\frac{1}{2}}F(x)[G(x)-G(\alpha)]^{\frac{1}{2}}+2G(x) \quad (3.9)$$

will be a lower bound for A whenever $\Delta_k^2(0)>2G(\alpha)$, with $A_{Lk}>A_{L1}$.

IV. UPPER BOUNDS FOR THE AMPLITUDE

It is assumed in this section that $u(x)$ is a monotone increasing function of x|| when $x<-x_0$; therefore, $v_p(x)\leq u(x)$ for $-A\leq x<-x_0$.

Let $x=\delta<-x_0$ be the abscissa of a point on $v=u(x)$. The contour $\lambda(x, v)=[u^2(\delta)/2]+G(\delta)$ thus, has at least one point in common with $v=u(x)$, when $x<-x_0$. A positive constant, k, can always be found satisfying $G(k)=[u^2(\delta)/2]+G(\delta)$. $\lambda_p(x, v)$ is strictly decreasing in $\delta\leq x<-x_0$. Therefore,

$$v_p(-x_0)<2^{\frac{1}{2}}[G(k)-G(x_0)]^{\frac{1}{2}}. \quad (4.1)$$

Now,

$$v_p(x)>2^{\frac{1}{2}}[G(x_0)-G(x)]^{\frac{1}{2}}, \quad (4.2)$$

when $-x_0\leq x\leq 0$. Using inequalities (4.1) and (4.2) and integrating Eq. (1.3), one obtains

$$v_p(0)<F(-x_0)+2^{\frac{1}{2}}[G(k)-G(x_0)]^{\frac{1}{2}}+[2G(x_0)]^{\frac{1}{2}}=v_{Uk}(0). \quad (4.3)$$

Since $y'\leq 0$ when y is above $F(x)$, by Eq. (1.7), then in the first quadrant of the Lienard plane $y_p(0)=v_p(0)>F(A)$. Therefore, $F(A)<v_{Uk}(0)$.

An upper bound for A is, then, the positive real root, $A_{Uk}>\alpha$, of the equation

$$F(x)=F(-x_0)+2^{\frac{1}{2}}[G(k)-G(x_0)]^{\frac{1}{2}}+[2G(x_0)]^{\frac{1}{2}}. \quad (4.4)$$

As $F(x)$ is positive and strictly increasing when $x>\alpha$, just one real root of Eq. (4.4) will exist.

The smallest value of A_{Uk}, namely A_{UK}, will occur when $k=K$ is the smallest k for which the contours $\lambda(x, v)=G(k)$ and $v=u(x)$ have at least one point in common.

V. SOME EXAMPLES

If $g(x)=x^{2n-1}$, $f(x)=\mu(x^{2n}-1)$, $\mu>0$, and n a positive integer, then $u(x)$ is strictly increasing for all $x<-x_0$ and every n. $G(x)=x^{2n}/2n$, $F(x)=\mu x[x^{2n}/(2n+1)-1]$. $x_0=1$, $\alpha=(2n+1)^{1/2n}$, $G(\alpha)=(2n+1)/2n$, $F(-x_0)=\mu[2n/(2n+1)]$. $F^2(x_0)>2G(\alpha)$ whenever $\mu^2>2[(2n+1)/2n]^3$ or, if $\mu>\frac{3}{2}3^{\frac{1}{2}}$ then $A_{L1}>\alpha$ will exist for every n. [A_{LK} will exist even when $\mu^2<2[(2n+1)/2n]^3$.

When $n=1$, Eq. (1.1), and $\mu=4$, $K=1.34^-$, $A_{LK}\doteq A_{L1}=1.85$, $A_{Uk}=2.14$; when $\mu=10$, $K=1.20^+$, $A_{L1}=1.96^+$, $A_{Uk}=2.05^+$. $\lim_{\mu\to\infty}A_{L1}=\lim_{\mu\to\infty}A_{LK}=\lim_{\mu\to\infty}A_{UK}=2$, a well-known result.[3, 5]

The author thanks Dr. A. T. Nordsieck for the differential analyzer solution shown in Fig. 1.

|| It would be sufficient to assume $v_p(x)\leq u(x)$ for $-A\leq x<-x_0$.

A Theory of the Envelope Type of Thermal Conductivity Tests

ARTHUR L. LOEB
Department of Metallurgy, Division of Ceramics, Massachusetts Institute of Technology, Cambridge, Massachusetts*

(Received July 12, 1950)

The well-known problem of the end effects in a cylinder of finite length has been circumvented by the use of spheroidal rather than cylindrical apparatus. After a general theory of heat flow is given, applications demonstrate the mathematical difficulties caused by discontinuities in the curvature of the surface of the heat conductor. The suitability of the spheroidal shape is demonstrated, and corresponding heat flow equations are derived demonstrating the effects of the variable distance between any two isotherms on the amount of heat crossing any unit area of an isotherm. It is proven that the amount of heat flowing through a slice of a spheroidal isotherm made perpendicular to the axis of the spheroid is proportional to the width of the slice, and independent of the position of the slice with respect to the spheroid. This theorem is applied to find the heat distribution necessary to keep the core of the apparatus at uniform temperature. Because all measurements are made at thermal equilibrium, a rough estimate is given for the time necessary to reach equilibrium.

INTRODUCTION

ONE idealized experiment for the measurement of thermal conductivity of ceramic materials at a high temperature consists of surrounding an infinitely long, uniformly heated cylindrical core by a hollow cylinder of the ceramic, and using the cylindrical form of the heat flow equation in determining the conductivity from the average temperature gradient across a certain distance in the ceramic.

The greatest deviation from idealized conditions is due to the finite length of the cylinder; and the present paper aims to justify the choice of an envelope in which the location of the isotherms, as well as the heat distribution necessary to keep the core heated uniformly, can be controlled. Whereas initially it was thought that hemispherical guards at the ends of the cylinder could be designed in such a manner that all lateral flow in the cylinder could be stopped, the analysis given below demonstrates that no such guards could eliminate nonradial flow entirely, while most heat will flow out per unit area through the guards, so that special precautions would have to be taken to prevent their cooling. It was therefore decided to streamline the apparatus into a spheroid, for which heat flow equations are derived below.

TABLE OF SYMBOLS

A = area of isothermal surface.
a = semifocal length of spheroidal coordinates.
subs a, b, refer to ceramic and insulating layers, respectively.
subs 1, 2, 3 refer to the surfaces separating core, ceramic layer, brick layer, and surroundings, respectively.
h^2 = thermal diffusivity.
k = thermal conductivity.
$n(z)$ = number of windings around core per unit axial length.
$\partial Q/\partial t$ = total amount of heat crossing isotherm per unit time.
r = radial coordinate.
t = time.
u = isothermal coordinate.
v = coordinate of flow.
W = heat crossing surface per unit area per unit time.
$s(u)$ = steady-state temperature function before heating core.
$T(u, t)$ = transient.
z = axial coordinate.
θ = temperature.
θ_c = final temperature of core.
ξ = const.: equation of spheroid.
η = const.: equation of hyperboloid of revolution.
φ = angle about axis of symmetry.
ψ = azimuthal angle.
Λ = solution of Laplace's equation.

A. GENERAL THEORY

In the envelope type of thermal conductivity tests a temperature gradient is maintained across the material to be measured by construction of a closed vessel with axial symmetry, the inside of which is heated uniformly, while the outside is kept at uniform room temperature. The core may be heated either by fitting a metal susceptor inside and placing the whole vessel in an inductor coil, or by wrapping platinum wire about the core, heating it by means of the heat developed in the platinum resistance.

The material to be measured may be surrounded by an insulating layer of brick in order to allow higher mean temperatures in the ceramic layer. The conductivity of the ceramic material is found by measuring the temperature gradient set up in it by a known quantity of heat flowing across it.

The shape of the core is arbitrary, and is chosen to fit experimental convenience. The heat flow should be perpendicular to surfaces with a discontinuity in conductivity, so that the surfaces $u=u_2$ and $u=u_3$ in Fig. 1 are determined by the thickness of the layers "a" and "b" as well as the shape of the core. Therefore, we use

* With funds from the AEC, contract AT(30-1)-602.

orthogonal coordinates $u=$const. (isotherms), $v=$const. (surfaces of flow), and $\varphi=$const. (planes through the axis of symmetry).

All observations are made at thermal equilibrium; the settling time to reach equilibrium has been found to be relatively independent of the shape of the vessel, and may be found for special simple cases.

Therefore, we proceed to solve the laplace equation subject to boundary conditions of constant core and environment temperatures, and continuous heat flow in the walls of the vessel:

$$\nabla^2\theta=0. \tag{1}$$

When $u=u_1$, $\theta=\theta_1$.

When $u=u_2$, θ and $k\partial\theta/\partial u$ continuous ($k\equiv$thermal conductivity).

If $\theta=\Lambda(u)$ is a solution of Eq. (1), then the solution satisfying the boundary conditions is:

$$\theta=\theta_1+\frac{\bar{k}}{k_a}\frac{\Lambda-\Lambda_1}{\Lambda_3-\Lambda_1}(\theta_3-\theta_1) \text{ in region "a"} \tag{2a}$$

$$=\theta_3+\frac{k}{k_b}\frac{\Lambda-\Lambda_3}{\Lambda_1-\Lambda_3}(\theta_1-\theta_3) \text{ in region "b,"} \tag{2b}$$

where $\Lambda\equiv\Lambda(u)$; $\Lambda_1\equiv\Lambda(u_1)$; $\Lambda_3\equiv\Lambda(u_3)$, $k_{a,b}\equiv$conductivity of region "a, b,"

$$(\Lambda_3-\Lambda_1)/\bar{k}=[(\Lambda_3-\Lambda_2)/k_b]+[(\Lambda_2-\Lambda_1)/k_a].$$

The amount of heat crossing an isotherm per unit time per unit area is

$$W\equiv-k\nabla\theta=-(\theta_3-\theta_1)\bar{k}\nabla\Lambda/(\Lambda_3-\Lambda_1).$$

N.B. Though Λ depends on u only, $\nabla\Lambda$ is a function of u as well as of v.

The amount of heat crossing an entire isotherm per unit time is:

$$\partial Q/\partial t=-(\theta_3-\theta_1)\bar{k}\oint\oint\nabla\Lambda\cdot dA/(\Lambda_3-\Lambda_1),$$

where

$$\oint\oint\nabla\Lambda\cdot dA$$

indicates integration over the entire isotherm (over-all values of v and φ); the expression is independent of u.

If we denote the difference in $k\theta$ between two points in either layer by $\Delta(k\theta)$, and the difference in Λ between

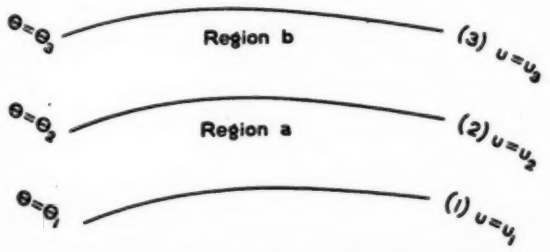


FIG. 1. Diagram of any double-layered wall.

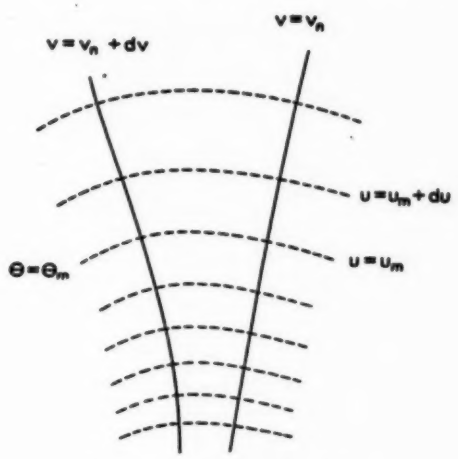


FIG. 2. Sketch of a tube of flow.

those points by $\Delta\Lambda$, we find:

$$\frac{\Delta(k\theta)}{\partial Q/\partial t}=-\frac{\Delta\Lambda}{\oint\oint\nabla\Lambda dA} \tag{3}$$

$$\Delta(k_a\theta_a)/\Delta(k_b\theta_b)=\Delta\Lambda_a/\Delta\Lambda_b. \tag{4}$$

In Eq. (3) the right-hand side is found from the geometry of the vessel, so that measuring $\Delta\theta$ and $\partial Q/\partial t$ gives k, whereas Eq. (4) may be used to check the result found for k_a when k_b is known, by measuring $\Delta\theta_b$ as well as $\Delta\theta_a$.

The amount of heat crossing an isotherm per unit time is given above. The amount of heat crossing a plane slab of comparable thickness across the same difference in temperature is proportional to the area of such a slab. In order to find out how large a slab would correspond to a vessel of comparable thickness, we equate the rate of heat transfer for the vessel and the slab, finding for the area of the slab:

$$\bar{A}=(u_3-u_1)\oint\oint\nabla\Lambda dA/(\Lambda_3-\Lambda_1). \tag{5}$$

There is always an isotherm with just that area; the area being dependent on u, we can then find the location of this "average isotherm."

$\nabla\theta$ is inversely proportional to the separation of the isotherms. Since heat flow is trapped between surfaces of flow, $\nabla\theta dA$ must be constant between any two flow surfaces. Therefore, near regions of high isotherm proximity, the flow converges (see Fig. 2).

Isotherms always crowd together near highly curved portions, so that we conclude that the lines of flow are also quite dense near such highly curved regions of the isotherms. Therefore, we expect that most heat leaves the core where the core is highly curved.

B. CHOICE OF THE SHAPE OF THE CORE

Experimental conditions require an elongated core, possibly a cylinder. In order to minimize heat flow from

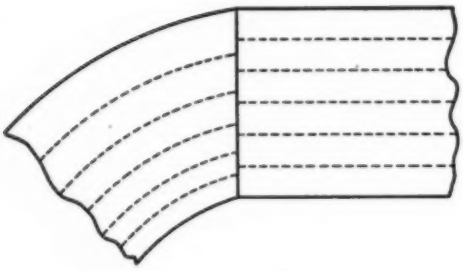


FIG. 3. Joint causing a discontinuity in the curvature.

the ends, hemispherical guards were installed. Analysis of the heat flow in spheres and in cylinders shows that it is impossible to have isotherms that are described by cylindrical surfaces joined along a plane to hemispherical surfaces; any different shape of the isotherms, however, causes heat flow across some portion of the plane of junction, so that no guards eliminate nonradial flow entirely. Therefore, a spheroidal core was designed, and the surfaces $u=u_1$, u_2, and u_3, together with the flow surfaces $v=$const., were found to be confocal spheroids and hyperboloids of revolution of two sheets.

C. CYLINDRICAL ISOTHERMS: ($u=r$, $v=z$)

$$\nabla_r^2=(1/r)(d/dr)(rd/dr).$$
$$\Lambda(r)=\log r.$$
$$\oint\oint\nabla\Lambda dA=2\pi l \quad (l=\text{length of the cylinder}).$$

Radius of the "average" isotherm:

$$\bar{r}=\frac{r_3-r_1}{\log(r_3/r_1)}=\frac{r_3+r_1}{2}\times\left\{\frac{1}{1+\frac{1}{3}(r_3-r_1)(r_3+r_1)+\frac{1}{5}[(r_3-r_1)/r_3+r_1]^3+\cdots}\right\}.$$

The radius of the average isotherm approaches the arithmetic mean between inside and outside radii of the cylinder from below as the thickness of the cylinder becomes small compared to the radius.

D. SPHERICAL ISOTHERMS: ($u=r$, $v=\psi$)

$$\nabla_r^2=(1/r^2)(d/dr)(r^2d/dr)$$
$$\Lambda(r)=1/r$$
$$\oint\oint\nabla\Lambda dA=4\pi$$
$$\bar{r}=\frac{r_1+r_3}{2}\left[1-\frac{1}{2}\frac{(r_3-r_1)^2}{(r_3+r_1)^2}+\cdots\right].$$

Therefore, the average isotherm's position obeys the same qualitative statement as that of the cylinder.

E. JOINED CYLINDER AND HEMISPHERES

Figure 3 shows how the isotherms crowd toward the core in the curved hemispherical portion of the wall. Although it is possible to vary the scale of this part of the drawing by varying the average conductivity of the guards, it is not possible to match up the hyperbolic temperature distribution in the guards with the logarithmic function in the cylinder. Generally, when two regions are joined such that their joint surface contains a discontinuity, then heat flow somewhere along their plane of junction causes distortion of their isotherms to join smoothly.

F. PROLATE SPHEROIDAL ISOTHERMS: ($u=\xi$, $u=\eta$)

$$\left.\begin{aligned}x&=a(\xi^2-1)^{\frac{1}{2}}(1-\eta^2)^{\frac{1}{2}}\cos\varphi\\ y&=a(\xi^2-1)^{\frac{1}{2}}(1-\eta^2)^{\frac{1}{2}}\sin\varphi\end{aligned}\right\}r=a(\xi^2-1)^{\frac{1}{2}}(1-\eta^2)^{\frac{1}{2}}$$
$$z=a\xi\eta$$
$$(r^2/\xi^2-1)+z^2/\xi^2=a^2;$$

$\xi=$const. describes prolate spheroids.

$$-(r^2/1-\eta^2)+z^2/\eta^2=a^2;$$

$\eta=$const. describes hyperboloids of revolution of two sheets.

a represents the semifocal length of isotherms and flow surfaces; if d_1 and d_2 are the distances from the foci to a point $P(\xi,\eta)$, ξ and η are given by

$$\xi=(d_1+d_2)/2a; \quad \eta=(d_1-d_2)/2a.$$

The value of a is found from the major and minor axes of the core.

$$\nabla^2=\frac{1}{a^2(\xi^2-\eta^2)}\left[\frac{\partial}{\partial\xi}(\xi^2-1)\frac{\partial}{\partial\xi}+\frac{\partial}{\partial\eta}(1-\eta^2)\frac{\partial}{\partial\eta}+\frac{\xi^2-\eta^2}{(\xi^2-1)(1-\eta^2)}\frac{\partial^2}{\partial\varphi^2}\right]$$
$$\Lambda(\xi)=\ln\frac{\xi-1}{\xi+1};\ 1<\xi<\infty;$$

$\Lambda(\xi)$ approaches for large values of ξ the solution for spherical isotherms, for $\xi\to1$ the solution for cylindrical isotherms.

To express the distance between isotherms $\xi=$const. and flow surfaces $\eta=$const., use

$$(ds)^2=(dr)^2+(dz)^2$$

Therefore,

$$\left(\frac{\partial s}{\partial\xi}\right)_\eta=a\left(\frac{\xi^2-\eta^2}{\xi^2-1}\right)^{\frac{1}{2}}; \quad \left(\frac{\partial s}{\partial\eta}\right)_\xi=a\left(\frac{\xi^2-\eta^2}{1-\eta^2}\right)^{\frac{1}{2}}.$$

Therefore,

$$\nabla\Lambda(\xi)=\frac{(\partial\Lambda/\partial\xi)_\eta}{(\partial s/\partial\xi)_\eta}=\frac{2}{a(\xi^2-1)^{\frac{1}{2}}(\xi^2-\eta^2)^{\frac{1}{2}}}.$$
$$dA=2\pi a^2(\xi^2-1)^{\frac{1}{2}}(\xi^2-\eta^2)^{\frac{1}{2}}d\eta$$
$$\oint\oint\nabla\Lambda dA=8\pi a.$$

Since $\nabla\Lambda dA=4\pi a d\eta$, and for a spheroid $\xi=$const.,

$$d\eta=dz/a\xi, \quad \text{where} \quad -a\xi<z<a\xi,$$

we can derive the following theorem.

When heat flows out perpendicularly through a spheroidal isotherm, the amount flowing through any slice made perpendicular to the axis of symmetry is directly proportional to the width of the slice, independent of the position of the slice with respect to the spheroid.

G. RESISTANCE HEATING OF THE CORE

The theorem stated above shows that, in order to keep the core heated uniformly, the amount of heat supplied must be the same for any unit length of the core. Since the amount of heat generated in the wire wrapped around the core is proportional to the length of wire, we must distribute the winding density suitably. Assuming that each turn of winding is a circle, the length of wire used per unit axial length is given by:

$$dL/dz = 2\pi r n(z),$$

where $n(z)$ = the number of windings per unit axial length.

If dL/dz is to be constant, $n(z) \propto 1/r$. It is impractical to use such a variable distribution; instead, the core is subdivided lengthwise into an odd number of equally wide sections, each section being wound uniformly by the same length of wire. The number of windings of the ith section is then given by:

$$n_i / \int_i r dz = 1/A_{ci},$$

where A_{ci} is the area of the lengthwise cross section of the ith slice.

Hence, the amount of windings for each section is given by

$$n_1 \div n_2 \div n_3 \div \cdots \div n_i \div \cdots = \frac{1}{A_{c1}} \div \frac{1}{A_{c2}} \div \frac{1}{A_{c3}} \div \cdots \frac{1}{A_{ci}} \cdots.$$

H. INDUCTANCE HEATING OF THE CORE

This will be discussed in a subsequent paper.

I. TIME TO SETTLE TO EQUILIBRIUM

The core is heated fairly rapidly, and consequently thermal equilibrium is disturbed. The time that elapses between disturbance and re-establishment of thermal equilibrium is made up of two terms: the time needed to heat up the core and the time needed in addition until thermal equilibrium exists again in the walls. Since infinitely slow heating does not disturb equilibrium, while instantaneous heating produces the worst disturbance, we can get a fair estimate of the settling time by assuming instantaneous heating, thus somewhat underestimating the time lag for heating the core, but overestimating the time needed for thermal equilibrium to re-establish itself.

The temperature obeys the following time-dependent equation:

$$\nabla^2\theta = (1/h^2)\partial\theta/\partial t, \quad \text{where } h^2 = \text{thermal diffusivity} \quad (6)$$

subject to the boundary conditions:

$$t=0: \quad \theta(u, 0) = s(u) \quad \text{for } u > u_1,$$

where $s(u)$ = steady-state temp. before heating.

$$t=0: \quad \theta(u_1, t) = \theta_c,$$

where θ_c is the final temperature of the core.

$$\theta(u_3, t) = s(u_3).$$

The solution is given by:

$$\theta(u, t) = s(u) + [(\Lambda_3 - \Lambda)/(\Lambda_3 - \Lambda_1)](\theta_c - \theta_1) + T(u, t), \quad (7)$$

where $T(u, t)$ satisfies Eq. (6), and vanishes at $t \to \infty$ (T represents transients), while

$$T(u, 0) = -(\Lambda_3 - \Lambda)(\theta_c - \theta_1)/(\Lambda_3 - \Lambda_1).$$

It is difficult to solve Eq. 6 for spheroidal coordinates, since in the nonsteady state there are no spheroidal isotherms in the walls. However, the settling time may be approximated by finding it for a spherical wall and for a plane slab. These two results agree so well that the same answer may be assumed for the spheroid. For the sphere

$$T(r, t) = \frac{2r_1}{r}(\theta_c - \theta_1) \times \sum_{n=1} \frac{(-1)^n}{n} \sin\left(n\pi \frac{r_3 - r}{r_2 - r_1}\right) \exp\left[-\frac{n^2\pi^2 h^2}{(r_3 - r_1)^2} t\right].$$

The settling time is usually defined as the time necessary for the transient to reach 1 percent of its initial value, and is seen to be proportional to the square of the thickness, and inversely proportional to the thermal diffusivity. For a double wall, we consider the two layers as being heated in series; the diffusivity being much greater for the inner layer and its thickness less, we regard only the settling time for the insulating outer layer as finite. The series for $T(r, t)$ converges so rapidly when the exponential term becomes of the order of 1 percent, that we drop all but the first term. The coefficient of the exponential term varies with the shape of the vessel, but this does not influence the settling time. It is seen that the spheroidal wall would settle to equilibrium most rapidly at the thinnest section, which causes heat to flow laterally until equilibrium is established throughout. The settling time is thus approximated by:

$$t_s = (2 \ln 10) d^2/\pi^2 h^2,$$

where d = greatest thickness of wall layer with lowest diffusivity, h^2 = lowest diffusivity.

ACKNOWLEDGMENT

The research for this article was done as a consulting assignment on the project conducted by Professor F. H. Norton with the assistance of Messrs. Milton Adams and Malcolm McQuarrie, whose original experiments were the inspiration for this work.

Long-Range Order in Beta-Brass and Cu_3Au

D. T. KEATING AND B. E. WARREN
*Massachusetts Institute of Technology, Cambridge, Massachusetts**
(Received November 9, 1950)

Measurements have been made on single crystals of CuZn and Cu_3Au held at various temperatures below T_c to decide whether there is a single-ordered phase whose order changes with temperature, or whether there is an ordered and a disordered phase the relative amounts changing with temperature. Precision measurements were made of the position and width of the (222) CuZn and (400) Cu_3Au reflections. The disordered phase has a larger cell dimension, so that reflections from the ordered and disordered phases are slightly displaced from one another. For both materials, a single peak was observed which was too narrow to correspond to the existence of two phases. At the stoichiometric compositions, it is concluded that below T_c there is only an ordered phase whose long-range order varies with temperature. Measurements of the long-range order in Cu_3Au have been made on briquets of filings, using quenched samples, and on a sample held at temperature.

INTRODUCTION

IN the beta-region of the Cu-Zn system above a critical temperature, $T_c=465°C$, the structure is disordered body-centered cubic, and below the critical temperature there is an ordered CsCl structure to which has been ascribed a varying degree of long-range order.[1] The copper-gold system at the composition Cu_3Au exhibits a similar phenomenon. The structure above the critical temperature of 394°C is disordered face-centered cubic, while below the critical temperature there is an ordered structure to which a varying degree of long-range order has been ascribed.[2,3]

However, the evidence of these investigators has been the x-ray integrated intensity measurements of superstructure reflections as an indication of the degree of long-range order. These measurements can be interpreted in an alternative manner; for example, as indicating the amount of the ordered phase present in an equilibrium mixture of ordered and disordered phases. From this standpoint, the square of the Bragg and Williams long-range order parameter is simply a factor representing the fraction of the material in the ordered state. Before the correct interpretation can be made of x-ray measurements, it must be established that there exists only one phase in these materials below the critical temperature.

Dilatometer measurements on beta-brass[4] and Cu_3Au[5] have shown anomalous increase in the thermal expansion of these alloys at temperatures somewhat below T_c. This is explained either by an increase of cell size with a decrease of order, or, if two phases exist, by an increase in cell size in changing from the ordered to the disordered phase. Samples of Cu_3Au quenched from the disordered phase exhibit a significantly larger cell size than those from the ordered phase.[6] Accordingly, the fundamental reflections from an ordered and disordered phase in equilibrium would be slightly displaced from each other. The relative intensities of the two reflections would vary as the relative amounts of the two phases present. A study of the fundamental reflections of these alloys was made at large glancing angles to resolve any splitting or to detect any anomalous broadening.

A STUDY OF THE TRANSITION IN BETA-BRASS

A highly ordered single crystal of beta-brass, which had been annealed at 400°C and cooled to room temperature slowly over a period of 33 hours, was mounted behind a nickel foil window in an electric furnace on a Geiger counter spectrometer. A steady flow of nitrogen through the furnace was maintained to prevent oxidation, and a thermocouple embedded in the crystal indicated its temperature. The (222) fundamental reflection was measured using Cu$K\alpha$ radiation focused on the entrance slit of the spectrometer by a bent quartz monochromator. By placing the chamber on a common circle with the entrance slit, the reflected radiation was focused approximately at the chamber. The horizontal divergence of the incident and reflected beams was as small as intensity considerations would permit. The crystal was turned uniformly through the Bragg reflection, and the total number of counts for various 2θ positions of the chamber recorded for each traverse of the sample. In this way, the shape of the (222) reflection was plotted for a series of temperatures ranging from room temperature to 550°C. Table I lists the time allowed before making any measurements for an equilibrium condition to develop. These times were

TABLE I. Times allowed for an equilibrium condition to develop in beta-brass.

Temp (°C)	300	350	400	420	440	460	480
Time (hr)	0.5	1.5	1.0 and 12.0*	0.5 and 12.0*	0.5 and 1.5*	0.5	0.5

* Measurements were repeated for these temperatures at times shown.

* Research sponsored by the ONR under contract N5ori07832.
[1] D. Chipman and B. E. Warren, J. Appl. Phys. **21**, 696 (1950).
[2] Z. W. Wilchinsky, J. Appl. Phys. **15**, 806 (1944).
[3] J. M. Cowley, J. Appl. Phys. **21**, 24 (1950).
[4] T. Matsuda, Sci. Repts. Tohoku Imp. Univ. XI, (1922).
[5] F. C. Nix and D. MacNair, Phys. Rev. **60**, 320 (1941).
[6] C. Sykes and H. Evans, J. Inst. Metals **58**, 255 (1936).

more than sufficient to bring the intensity of a superstructure line to equilibrium.[1]

Plots of the (222) reflections for the temperatures listed in Table I were all regular in shape; no splitting or broadening of the reflection was observed. The α_1, α_2 doublet was completely resolved, and the half-maximum peak widths averaged nearly 7.5 minutes in 2θ. The natural line width of copper $K\alpha_1$ is 5.4 minutes in 2θ at this angle.

Figure 1 shows the 2θ position of the (222) reflection as well as the peak widths as a function of temperature. The upper dashed line is an extrapolated position for the reflection from an ordered cell, the lower dashed line the extrapolated position for the reflection from a disordered cell. Since the separation of the extrapolated positions is 18 minutes and the peak widths only 7.5 minutes, the reflections resulting from two phases would have resolved into two peaks lying along the extrapolated positions. This was not the case, and the anomalous shift in peak position around 400°C is due to the increase in cell size accompanying the disordering in a single phase.

TABLE II. Times allowed for an equilibrium condition to develop in Cu_3Au.

Temp (°C)	252	300	322	340	360	380	390
Time (hr)	50.5	67.2	4.0 and 23.0*	6.0 and 48.5*	1.0 and 18.0*	22.0	17.5

* Measurements were repeated for these temperatures at times shown.

A STUDY OF THE TRANSITION IN Cu_3Au

A single crystal of composition close to Cu_3Au was first brought to a high degree of order by annealing it for 25 hr at 760°C, 50 hr at 400°C, 100 hr at 350°C, 70 hr at 300°C, 50 hr at 250°C, and 50 hr at 175°C.

The same experimental technique was used to study the (400) fundamental reflection of Cu_3Au using $CoK\alpha$ radiation. The crystal, being small, had to be set into a brass block into which the thermocouple was embedded, and a hydrogen atmosphere had to be used.

Measurements were taken on the (400) reflection after allowing the times, listed under the corresponding temperatures in Table II, for the sample to reach equilibrium. These times were sufficient for an equilibrium value of the intensity of a superstructure reflection to be reached (see Table IV).

Plots of the (400) reflection for the temperatures and times in Table II were all regular, and no splitting or anomalous broadening was observed. The half-maximum widths averaged nearly 18 minutes in 2θ. The natural line width of $CoK\alpha_1$ is 9.7 minutes at this angle. In Fig. 2 the 2θ position of the (400) reflection is plotted along with the half-maximum widths as a function of temperature. Again the upper dashed line is an extrapolated position for the reflection from an ordered cell, the lower from the disordered cell. The extrapolated

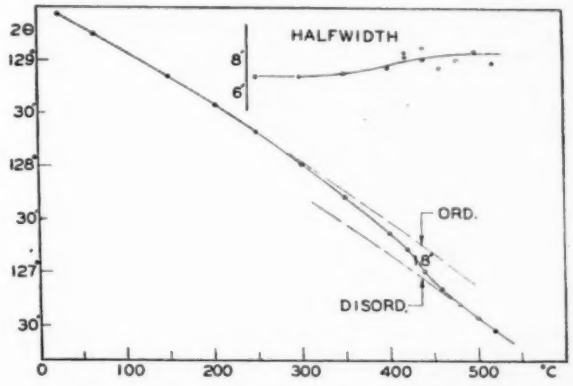


FIG. 1. Change with temperature of the line position and width for (222) of β-brass.

positions are separated by 30 minutes. Clearly, two phases would have resolved into two peaks lying along the extrapolated positions. Instead, an anomalous shift in peak position, due to the increasing disorder in a single phase, was observed. The discontinuous shift in peak position is due to the discontinuous decrease in long-range order at the critical temperature in Cu_3Au.

MEASUREMENT OF LONG-RANGE ORDER IN Cu_3Au

Experimental points for S, the Bragg and Williams long-range order parameter, have been too few[2,3] to give a reliable dependence of long-range order on temperature. Accordingly, an ingot of the stoichiometric composition Cu_3Au was prepared and given a homogenizing anneal of 200 hr at 920°C and 100 hr at 860°C. Filings were made from the ingot; and those filings, passing through a 325 mesh, were used to face the top $\frac{1}{32}$ in. of four $\frac{1}{2}$ in.$\times\frac{1}{2}$ in.$\times\frac{1}{8}$ in. pressed briquets made from the filings. The four briquets, 1, 2, 3, and 4, were sealed off in vacuum and annealed according to the times listed in Table III. Each sample was removed directly from the furnace and quenched in order. Thus, each succeeding sample was subjected to the history of the preceding sample.

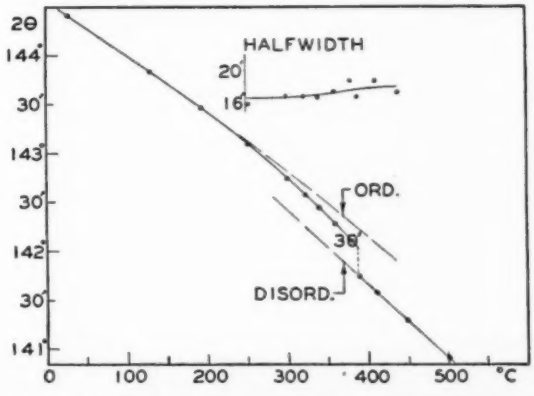


FIG. 2. Change with temperature of the line position and width for (400) of Cu_3Au.

TABLE III. Annealing times, temperatures, and values of S for quenched samples of Cu_3Au.

Sample	Temp. (°C)	Ann. time (hr)	P_{200}	P_{210}	P_{222}	S[a]
No. 1	381	56	914.0	116.8		0.585
No. 2	350	131	949.0	269.9		0.873[c]
No. 3	300	137	779.0	283.3		0.988[c]
No. 4	250	120				
	200	200				
	150	100	737.0	275.0	166.0	1.00[c]
No. 1[b]	375	71	864.4	220.9		0.830[c]
No. 2[b]	375	71	819.8	228.7		0.862[c]

[a] Atomic scattering factors corrected for dispersion: $\Delta f_{Cu} = -2.51$, and $\Delta f_{Au} = -6.0$. [R. W. James, *The Optical Principles of the Diffraction of X-Rays* (G. Bell and Sons, Ltd., London, 1948), pp. 608 and 187, respectively.]

[b] Samples No. 1 and No. 2 were reannealed after the first set of measurements, as it was felt that 0.585 did not represent the equilibrium value. (See references 6 and 7.)

[c] Plotted in Fig. 3.

Two methods were used for determining S on the quenched samples and agreed within about 5 percent. In all cases Cu$K\alpha$ radiation, monochromated by the $(10\bar{1}1)$ plane of a bent quartz crystal, illuminated an area 0.17 in.×0.375 in. on the sample, and again approximate focusing at the chamber was employed. The first method employed a narrow counter slit and amounted to plotting out the diffraction peaks and comparing areas. The second substituted a chamber slit wide enough to allow all the radiation from a diffraction peak to enter the chamber and compared the total number of counts corrected for background. The average of the number of counts on both sides of the peak was used as background. In both cases symmetrical reflection off the flat face was maintained, and all readings were corrected for half-wavelength radiation by a technique employing nickel and aluminum filters balanced at the half-wavelength.

The formula for S, from comparison of the integrated intensities of a superstructure and fundamental line, is:

$$S^2=\left(\frac{P_{\text{sup}}}{P_{\text{fund}}}\right)\frac{[mF^2e^{-2M}(1+\cos^2 2\theta_m\cos^2 2\theta)/\sin^2\theta\cos\theta]_{\text{fund}}}{[mF^2e^{-2M}(1+\cos^2 2\theta_m\cos^2 2\theta)/\sin^2\theta\cos\theta]_{\text{sup}}}. \tag{1.0}$$

$(P_{\text{sup}}/P_{\text{fund}})$ is the ratio of the observed intensities of a superstructure and fundamental line; m = multiplicity; F = appropriate structure factor; $M=8\pi^2(\langle u^2\rangle_{Av}/\Lambda^2)\sin^2\theta$, the temperature factor; $\langle u^2\rangle_{Av}$ = mean square atomic displacement; Λ = wavelength of x-rays; θ_m = Bragg angle of monochromator; and θ = Bragg angle of reflection.

$\langle u^2\rangle_{Av}/\Lambda^2$ is conveniently found by the comparison of the observed intensities of two fundamental lines, *hkl* and *lmn*.

$$\langle u^2\rangle_{Av}/\Lambda^2=\frac{-\ln\left\{\left(\frac{P_{hkl}}{P_{lmn}}\right)_{\text{obs}}\frac{[mF^2(1+\cos^2 2\theta_m\cos^2 2\theta)/\sin^2\theta\cos\theta]_{lmn}}{[mF^2(1+\cos^2 2\theta_m\cos^2 2\theta)/\sin^2\theta\cos\theta]_{hkl}}\right\}}{16\pi^2(\sin^2\theta_{hkl}-\sin^2\theta_{lmn})}. \tag{1.1}$$

A value of $\langle u^2\rangle_{Av}/\Lambda^2=0.00495$ for room temperature was obtained from sample No. 4 using the (200) and (222) reflections. Table III gives the values of S obtained from the quenched samples.

Secondary extinction is eliminated by using powders. Primary extinction, which is nevertheless present, is indeterminate depending upon the state of perfection of each small crystal. In general, primary extinction is greater in annealed samples and is more pronounced for strong reflections. An extinction coefficient defined as (I with extinction)/(I without extinction) is of use, and it was determined by comparing the (200) reflection of sample No. 3 with the (220) reflection of α-brass, a good intensity standard,[7] according to Eq. (1.2) when the incident radiation is unpolarized.

$$\text{Extinction coefficient}=\left(\frac{P_{\text{CuAu}_3}}{P_{\alpha\text{-brass}}}\right)\frac{[mF^2e^{-2M}(1+\cos^2 2\theta)/u_l v^2\sin^2\theta\cos\theta]_{\alpha\text{-brass}}}{[mF^2e^{-2M}(1+\cos^2 2\theta)/u_l v^2\sin^2\theta\cos\theta]_{Cu_3Au}} \tag{1.2}$$

u_l = effective absorption coefficient, and v = volume of the unit cell. The measured ratio of intensities, 5.68 in.^{2}/5.30 in.^{2}, yielded an extinction coefficient of 0.90.

It is reasonable to expect the (210) reflection to suffer extinction of lesser extent. Nevertheless, the error introduced into the ratio of the (200) and (210) intensities is less than the relative error occurring in either intensity as the result of extinction, and the relative error in S is correspondingly less.

To measure the value of S at temperature, sample No. 4 was selected and mounted in the furnace of the spectrometer. An aluminum foil window and a hydrogen atmosphere were used. The sample was heated to various temperatures and maintained at each temperature until the intensity of the (210) superstructure reflection approached an equilibrium value (see Table

[7] B. L. Averbach and B. E. Warren, J. Appl. Phys. **20**, 1066 (1949).

IV). If the sample has been annealed for a sufficient time to eliminate all the small antiphase domains, so that the superstructure reflections will not be so spread out as to confuse the wings of the reflection with the background, then the measured integrated intensity is independent of the existence of any remaining larger antiphase domains. When the crystal is in this highly ordered state, it is possible to insure that equilibrium has been reached at any temperature. If the integrated intensity of a superstructure line is measured at a particular temperature below T_c, the temperature is then lowered and held until a higher intensity characteristic of a higher order is reached, and the temperature then raised to the original value, *the intensity decreases to the original value.* The open slit method was used, and sample and chamber settings were corrected for shift in peak position with temperature.

If the long-range order parameter is S_0 at room temperature, it may be found at some elevated temperature by Eq. (1.3).

$$S=S_0\left\{\left(\frac{P_T}{P_0}\right)_{\text{sup}}\times\left(\frac{P_0}{P_T}\right)^{\gamma}_{\text{fund}}\right\}^{\frac{1}{2}}, \qquad (1.3)$$

where

$$\gamma=(h^2+k^2+l^2)_{\text{sup}}/(h^2+k^2+l^2)_{\text{fund}}$$

$(P_T/P_0)_{\text{sup}}$ = ratio of elevated and room temperature intensities of a superstructure reflection, and $(P_0/P_T)_{\text{fund}}$ = ratio of room and elevated intensities of a fundamental reflection, the (222) in this case. The value of S_0 was determined by comparison with a fundamental reflection.

Values of S are plotted in Fig. 3. The times for temperatures over 384°C in Table IV are only approximate. It was found that while disorder proceeded rapidly at 388°C, recovery was slow. For instance, after making the measurement at 390°, a measurement of S at 387° showed no appreciable recovery from its value at 390° after about 15 minutes. When the measurement of S was repeated at 390°, the original value was obtained. Values of S in the neighborhood of 390°–394°C must be considered as approximate and represent what is in effect an experimental measurement of a theoretically discontinuous change. The results lend strong support to the belief that order, once having been lost discontinuously, is a long time in being restored. A critical temperature of 394°C is probably good within ±3°C. In general, the agreement with Cowley's theoretical considerations is remarkably good.[8]

TABLE IV. Equilibrium times and values of S from sample No. 4 at elevated temperatures.

Temp. (°C)	Time (hr)	P_{210}	P_{222}	S
30	—	421.4	263.0	1.00*
250	24.	404.9	196.4	1.04
250	52.	358.3	204.4	0.972*
300	1.5	317.4	172.5	0.945
300	21.	311.3	172.5	0.944*
350	2.5	245.1	174.5	0.842
350	24.	253.0	177.5	0.841*
370	3.	197.4	131.5	0.744
370	5.	197.4	131.5	0.787*
380	.5	207.2	154.7	0.785
380	1.5	201.2	154.7	0.775*
384	0.2	207.0	—	0.785*
387	0.2	203.2	—	0.775*
388	0.2	159.0	—	0.686*
390	0.2	86.0	—	0.505*
392	0.2	30.5	—	0.300*
394	0.2	16.0	143.6	0.216*

* Plotted in Fig. 3.

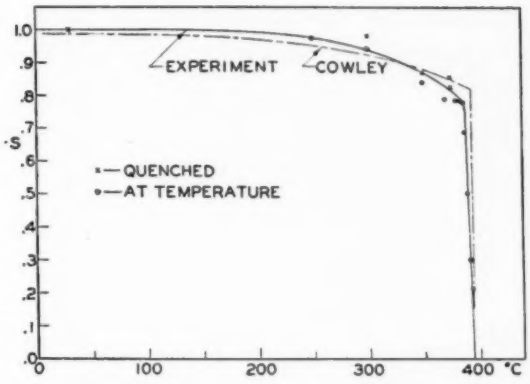


FIG. 3. Variation with temperature of the long-range order parameter for Cu_3Au.

CONCLUSIONS

In β-brass below T_c the variation in the intensity of a superstructure line is due to the variation in the long-range order parameter in a single-ordered phase, rather than to a change in the relative amounts of an ordered and a disordered phase. The sample temperature was uncertain by perhaps ±5°C; and, hence, these measurements do not exclude the existence of a very narrow two-phase region at the critical temperature. Below such a narrow temperature interval, there is only a single-ordered phase with a continuously varying long-range order parameter which is unity at low temperature and approaches zero at T_c. At each temperature below T_c, there is an equilibrium value for the long-range order parameter; the same value is obtained by approaching from either above or below.

For a sample close to the stoichiometric composition Cu_3Au which has been brought to equilibrium at a temperature below T_c, the measurements indicate that there is only a single-ordered phase with a long-range order parameter which varies from unity at low temperature to about $S=0.8$ at the critical temperature. Providing that the sample has been well ordered so that the antiphase domains are largely eliminated, there is an equilibrium value for S at each temperature below T_c, and the same value is obtained by approach from either above or below. The sample temperature

[8] J. M. Cowley, Phys. Rev. **77**, 669 (1950).

was uncertain by perhaps ±5°C; and, hence, these measurements do not exclude the existence of a narrow two-phase region at T_c. Since the composition was close to Cu_3Au, these measurements do not exclude the existence of a wider two-phase region at compositions well removed from Cu_3Au.

The measurements reported in this paper were on alloys which were in as nearly an equilibrium condition as it was possible to obtain in reasonable times. The results and conclusions have to do only with the equilibrium condition below T_c for β-brass and Cu_3Au; they have nothing to do with the mechanism or the kinetics of long-range ordering.[9,10]

[9] Newkirk, Geisler, Martin, and Smoluchowski, Trans. Am. Inst. Mining Met. Engrs., Inst. Metals Div. **188**, 1249 (1950).
[10] A. Guiniér and R. Griffoul, Rev. mét. **45**, 387 (1948).

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

Phase Equilibria in an Ordering Alloy System

J. B. NEWKIRK AND R. SMOLUCHOWSKI
Carnegie Institute of Technology, Pittsburgh, Pennsylvania

AND

A. H. GEISLER AND D. L. MARTIN
General Electric Research Laboratory, Schenectady, New York

(Received November 1, 1950)

The equilibria between phases in platinum alloys containing 42, 48, and 54 atomic percent cobalt were determined at temperatures below 1000°C. This system of alloys exhibits an order-disorder transition analogous to that for CuAu. A temperature range in which ordered and disordered phases coexist in equilibrium was determined for each alloy. These data were incorporated in the phase diagram. The Co-Pt system is the first ordering alloy system in which it has been shown that this reaction is a true phase transformation contrary to the prevalent concepts of the phenomena.

INTRODUCTION

A TENDENCY has developed in the last two decades to treat the order-disorder reaction in alloy systems as a homogeneous process typified by the so-called "phase transition of the second type." Little data were available which could be interpreted solely on the basis of a classical phase transformation according to Gibbs' Phase Rule. The fact, however, that all data could be reconciled with classical theory of phase transformations, when due allowance was made for limitations of experimental techniques and lack of thermodynamic equilibrium, prompted the authors to make an extensive investigation of the process in a system of ordering alloys. Brief reports of the research have been published in the past.[1,2] A detailed report of the kinetics of the reaction whereby disordered alloys were transformed isothermally to the single-phase ordered state is presented in another paper.[3] The portion of the research to be reported here concerns the equilibrium states attained by prolonged heating which insured completeness of the reaction.

Three alloys of the Co-Pt system containing 42, 48, and 54 atomic percent Co were prepared by melting and casting in an atmosphere of nitrogen. Microscopic examinations were made on specimens from a rod $\frac{1}{4}$ inch in diameter which had been hot swaged from the castings. x-ray diffraction analyses and electrical resistance measurements were made on 0.020-inch wire or 0.015 inch thick sheet (in the case of the 42 Co alloy), fabricated from the $\frac{1}{4}$-inch rod. The initial disordering treatment consisted of holding specimens sealed in quartz vials containing argon at low pressure, several hours at 900–1000°C. The disordered state was retained by quenching the vial into water. During subsequent thermal treatments the specimens were likewise protected from oxidation and quenched to insure retention of the structure characteristic of the high temperature. Additional details of specimen preparation are reported elsewhere.[3]

REPORTED PHASES

The most recent phase diagram of the Co-Pt system is one by Gebhardt and Köster[4] and is reproduced in Fig. 1. The salient features near the composition Co-Pt are, according to these investigators: (1) the disordered face-centered cubic solid solution γ_{CD}, at temperatures between 825°C and the solidus; (2) a disordered face-centered tetragonal phase γ_{TD}, between 825° and 520°C; and (3) an ordered face-centered tetragonal phase γ_{TO}, at temperatures below 520°C. Usually in face-centered cubic alloys of the AB-type a tetragonal lattice occurs concurrent with the ordering reaction; hence, the above findings of Gebhardt and Köster of a disordered tetragonal phase were questionable. Their evidence for this was an inflection in

[1] Newkirk, Geisler, and Martin, J. Appl. Phys. **20**, 816 (1949).
[2] J. B. Newkirk and R. Smoluchowski, Phys. Rev. **76**, 471 (1949).
[3] Newkirk, Geisler, Martin, and Smoluchowski, J. Metals **188**, 1950), (10), 1249–1260.
[4] E. Gebhardt and W. Köster, Z. Metallkunde **32**, 253–261 (1940).

the resistance-temperature curve which resembled the change reported for other order hardening alloys (e.g., CuAu) at the critical temperature of order; however, they were unable to obtain x-ray verification of γ_{TD} in quenched samples, which they interpreted by the claim that ordering could not be suppressed by quenching. Results of the present investigation indicate that only the ordered face-centered tetragonal phase, γ_{TO}, exists in equilibrium throughout the single phase field below 825°C. To prove this, high temperature Debye-Scherrer photographs were made on wire specimens of the 48 atomic percent cobalt alloy at 500°C, 600°C, and 720°C. The patterns resulting from all three of these tests contained all the characteristic lines of an ordered face-centered tetragonal structure. No diffraction lines were present in these high temperature photograms whose position and intensity could not be accounted for by the ordered structure.

A body-centered cubic phase has been reported by Jellinghaus in cast samples of Co-Pt,[5] and by Hultgren and Jaffee[6] in a powder sample held for about 500 hours at 400°C. In the present investigation, no evidence of this phase was found by Debye-Scherrer studies, even in specimens held nearly twice as long at 400°C as those used by Hultgren and Jaffee. However, it was found necessary to protect the specimens carefully from oxidation or other attack during annealing in order to avoid the appearance of spurious lines in x-ray patterns due to oxides or other surface impurities. Gebhardt and Köster show a two-phase field between γ_{CD} and γ_{TO} (Fig. 1), without direct evidence for it. Later Hultgren and Jaffee found that a powder sample which had been held for a long time at 800°C, gave a Debye-Scherrer photograph consisting of two patterns, one disordered face-centered cubic, γ_{CD}, and the other ordered face-centered tetragonal γ_{TO}. This structure which was the first evidence for a two-phase field has been confirmed in the present study.

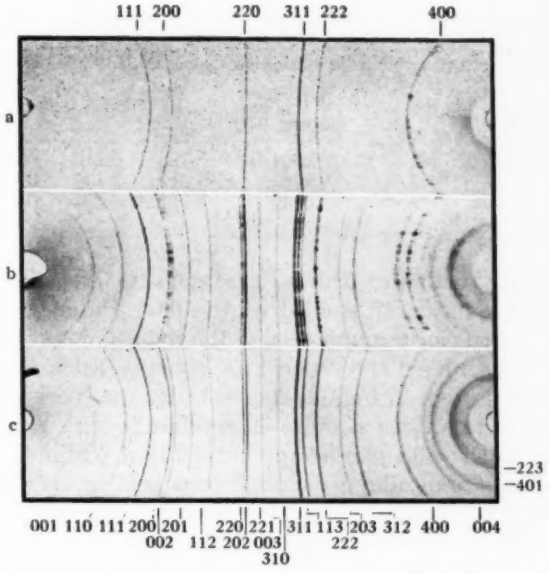


FIG. 2. Debye-Scherrer patterns illustrating three equilibrium structures of 54 atomic percent Co alloy. (a) Disordered face-centered cubic phase, γ_{CD}. (b) Disordered and ordered phases coexisting in equilibrium. (c) Ordered face-centered tetragonal phase, γ_{TO}.

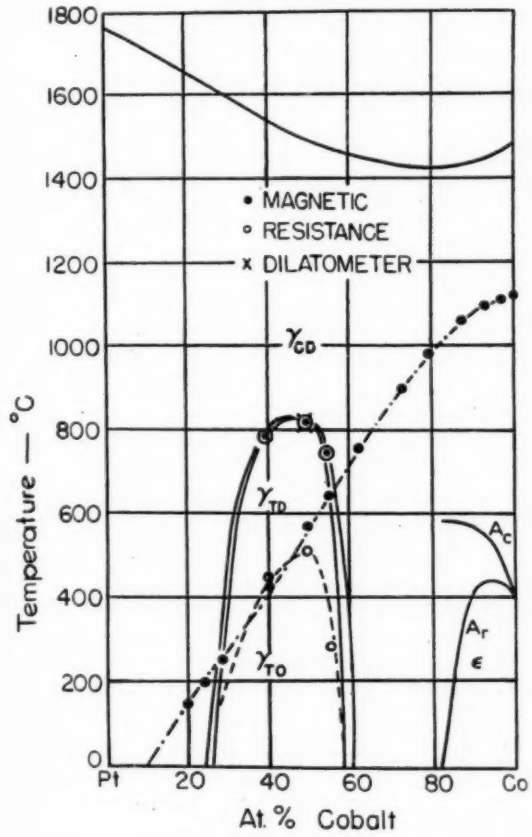


FIG. 1. Constitution diagram for the Co-Pt system according to Gebhardt and Köster.[4]

STRUCTURE AND COEXISTENCE OF THE PHASES

The diffraction photographs of Fig. 2a and Fig. 2c show the Debye-Scherrer patterns characteristic of the disordered, γ_{CD}, and the ordered, γ_{TO}, states respectively. The pattern in Fig. 2a shows that the disordered lattice is face-centered cubic. Comparison of calculated with observed line intensities in Fig. 2c indicates that the unit cell sketched in Fig. 3 is the correct one for the

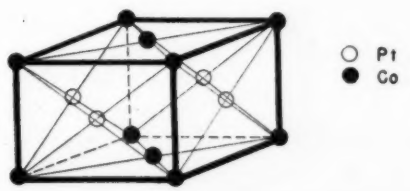


FIG. 3. Unit cell for the ordered phase.

[5] W. Jellinghaus, Z. tech. Physik **17**, 33 (1936).
[6] R. Hultgren and R. I. Jaffee, J. Appl. Phys. **12**, 501 (1941).

TABLE I. Lattice parameters for disordered and ordered Co-Pt alloys.

Weight % Co	Atomic % Co	Disordered 1 hr @ 1000°C, W.Q.	Ordered at 700°C		
			a	c	c/a
17.71	41.62	3.801A	3.798	3.746	0.987
21.61	47.64	3.782	3.812	3.708	0.973
26.01	53.80	3.749	3.771	3.694	0.980

ordered structure. This is analogous to the ordered phase in the much studied CuAu alloy. The measured dimensions for the unit cells of the three alloys studied when disordered and ordered are given in Table I.

The pattern in Fig. 2b demonstrates the coexistence of discrete regions of order in equilibrium with regions of disorder. This photograph is typical of specimens of the 54 percent alloy which had been held several hundred hours in the range of 720–810°C and quenched into water. Every line in this pattern can be matched both in position and in relative intensity with a line in either the photograph above it or below it. That the two-phase equilibrium can be attained from either ordered or disordered starting states was shown by the following experiment. An initially ordered wire of the 48 atomic percent cobalt alloy (96 hours at 700°C, air-cooled) was sealed in a quartz tube together with an initially disordered wire of the same composition (1 hour at 1000°C, W.Q.). The two wires were then held at a temperature within the two-phase region for over 100 hours, after which time the tube was quenched into water. Both wires were found to contain a mixture of ordered and disordered phases discretely separated. A parallel experiment, using wires of the 54 atomic percent cobalt alloy, gave similar results.

SURFACE EFFECT

It was found that the surface of drawn wires behaved somewhat differently from the interior which was left when a thin layer of the original surface was etched off with aqua regia. These differences in behavior were apparently independent of the stage of treatment at which the original outside layer was removed. The observed dissimilarities are summarized as follows:

1. The cubic lattice parameter at the surface of a drawn wire is slightly greater than that at the interior.
2. The grain size is much smaller at the surface.
3. The equilibrium transition from order to disorder takes place at a higher temperature at the surface. (Fig. 7a).
4. The two-phase range is wider for the surface metal.
5. The ordering rate is higher at the surface.

The finer grain size on the outside of unetched wires is probably a consequence of the uneven degree of plastic deformation which the metal received during the forming operations. The other differences may be due to the combined effects of the uneven grain size and the presence of a small amount of impurity such as carbon which might have been picked up by the surface metal while the rough bar was being fabricated. This difference in behavior of the inside and outside metal prolonged the investigation somewhat but did not alter the qualitative conclusions which could be drawn from it.

TABLE II. Comparison of equilibrium temperatures determined by various methods.

42 atomic percent Co	X-ray diffraction		Electrical resistivity			
	Etched	Unetched	at temp.	at 0°C	Micro. exam.	Average
$D \rightleftharpoons D+O$ boundary	785	—	—	786, 790	—	787
$D+O \rightleftharpoons O$ boundary	783	—	—	775, 784	—	781
Midpoint of range	784	—	—	780, 787	—	784
48 atomic percent Co						
$D \rightleftharpoons D+O$ boundary	824	833	818	832, 827	826	826
$D+O \rightleftharpoons O$ boundary	820	826	814	815, 814	—	818
Midpoint of range	822	830	816	824, 820	<823	822
54 atomic percent Co						
$D \rightleftharpoons D+O$ boundary	736	812	717	755, 737	>736	749
$D+O \rightleftharpoons O$ boundary	704	730	707	742, 700	—	717
Midpoint of range	720	771	712	749, 718	<736	734

LIMITS OF TWO-PHASE REGIONS BETWEEN ORDERED AND DISORDERED PHASES

X-Ray Diffraction Results

In order to locate the upper and lower temperature limits of the two-phase regions, the existence of which was demonstrated by Fig. 2b, wire specimens of the three alloys were held at high temperatures for long periods of time, water quenched, and then examined by the Debye-Scherrer method. Approximately 70 specimens were used to investigate different treatments. Some of these were initially ordered, whereas the bulk initially were in the disordered condition. Specimens were heated at increments of 2 to 3 degrees in the temperature range of the two-phase equilibrium. Holding times varied from 16 to 368 hours with an average time of 112 hours. Periods of 16 hours and less were insufficient to establish equilibrium at some of the temperatures. Many of the specimens were etched to remove the surface layer before making the patterns. The results were used to locate the phase boundaries which are listed in Table II under the columns "X-ray diffraction, etched and unetched." These will be discussed later when they are compared with results of other techniques.

Electrical Resistance Measurements

The variation of resistance with temperature was measured for the three alloys by heating and cooling a specimen sealed in a long quartz glass tube. The measurements were made with a Kelvin bridge by connecting to four nickel lead wires which extended through the sealed specimen tube. The general practice for heating and cooling was to hold the specimen at the set temperature and then to increase or decrease the temperature to the next step. Over week-ends the specimen would be at a temperature about 65 hours. The time used to obtain a complete heating and cooling curve

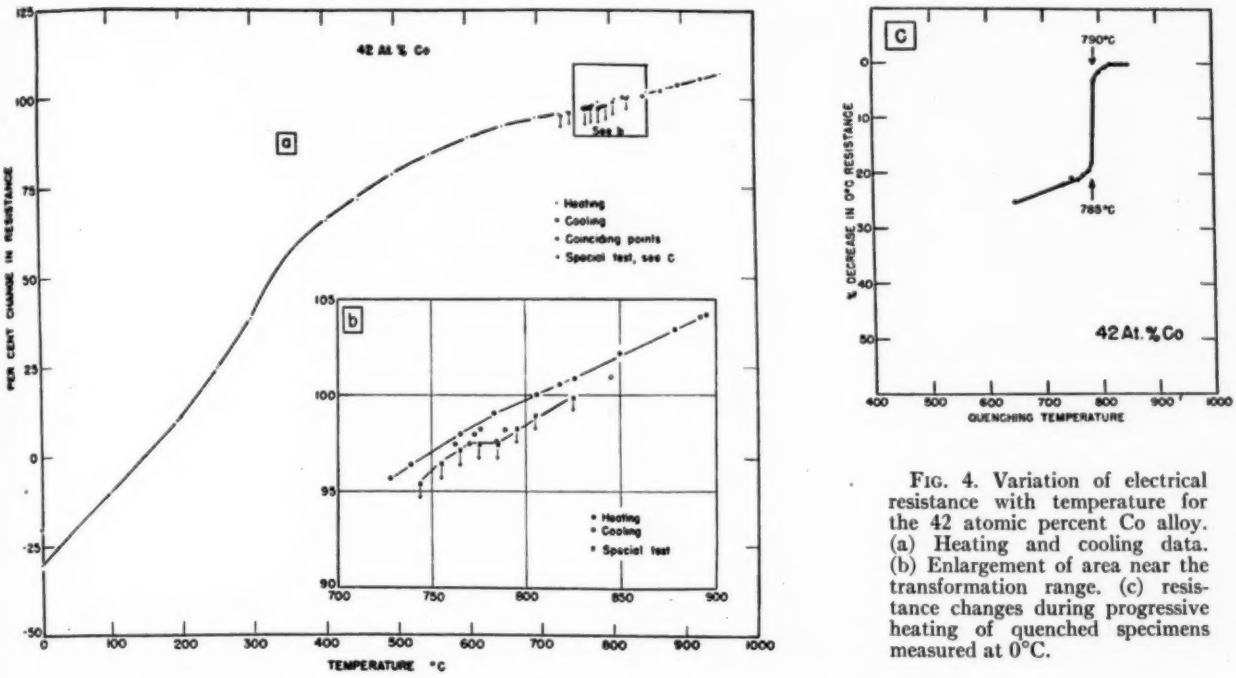


FIG. 4. Variation of electrical resistance with temperature for the 42 atomic percent Co alloy. (a) Heating and cooling data. (b) Enlargement of area near the transformation range. (c) resistance changes during progressive heating of quenched specimens measured at 0°C.

was generally more than two weeks. Near the critical temperature the steps were only a few degrees in magnitude.

The results of these heating and cooling studies on the three alloys are presented in Figs. 4–6. Since several specimens of varying lengths were used and resistivity measurements at high temperatures are difficult to make because of the uncertainty in specimen dimensions, it was decided to plot the percent change of resistance of the initially disordered alloy *vs* the temperature.* The agreement between different specimens and repeat runs is good. The inflections in the curves at about 400° for the 42 percent alloy, 475° for 48 percent, and 600° for the 54 percent alloy correspond to the magnetic transition (curie) temperature rather than a change in atomic arrangement as assumed by

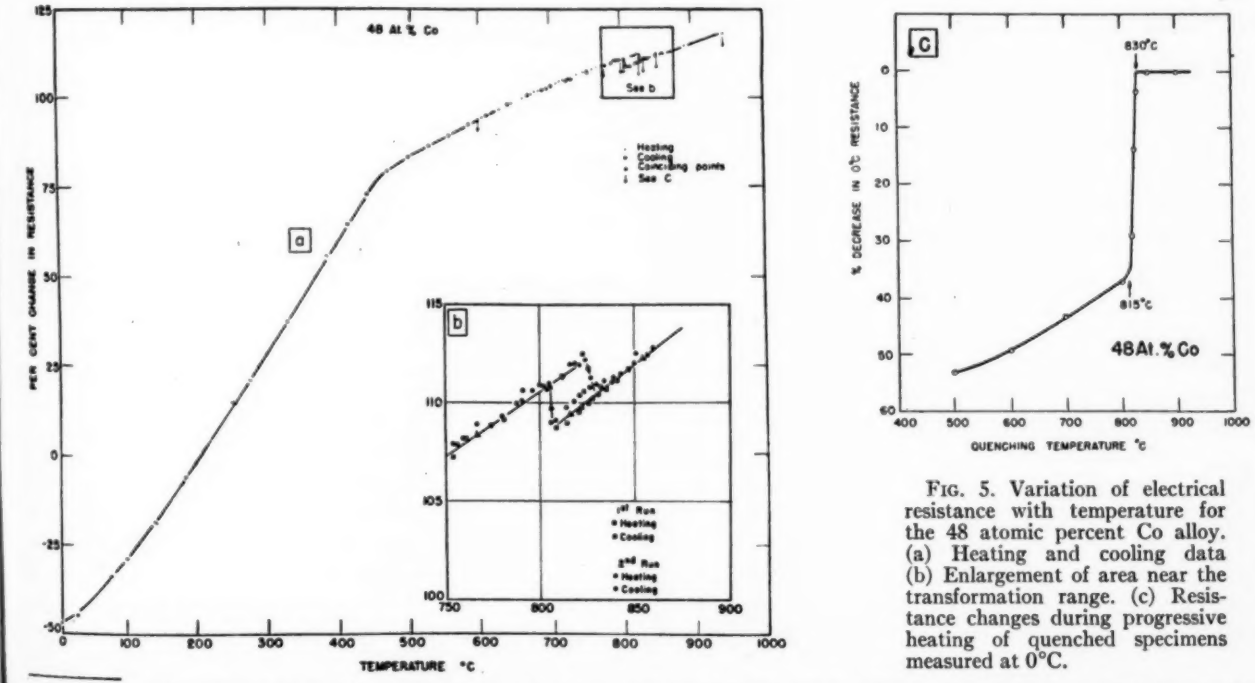
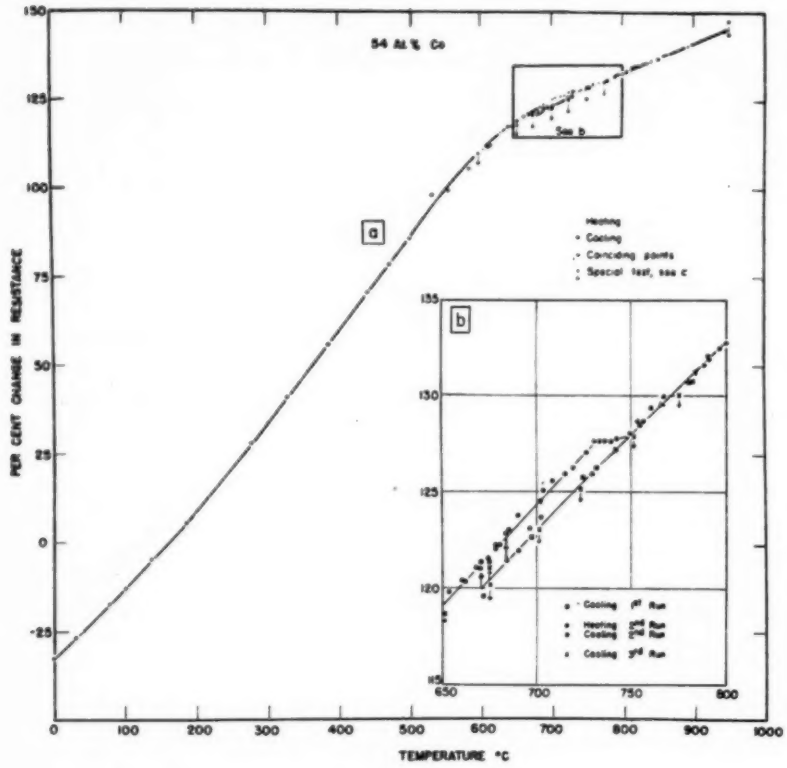


FIG. 5. Variation of electrical resistance with temperature for the 48 atomic percent Co alloy. (a) Heating and cooling data (b) Enlargement of area near the transformation range. (c) Resistance changes during progressive heating of quenched specimens measured at 0°C.

* The base resistance was that of the specimen quenched from the disordered region and measured at 0°C; when variations of this value for a given specimen were found, the highest value was generally used for determining the percent change.



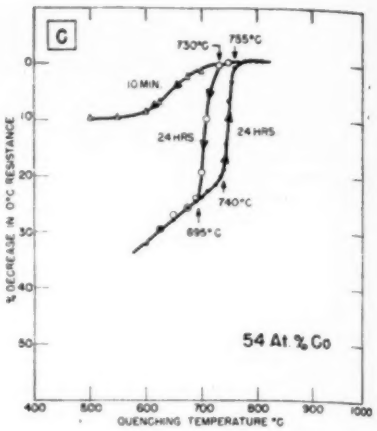


FIG. 6. Variation of electrical resistance with temperature for the 54 atomic percent Co alloy. (a) Heating and cooling data. (b) Enlargement of area near the transformation range. (c) Resistance changes during progressive heating and cooling of quenched specimens measured at 0°C.

Gebhardt and Köster. The illustrations, Figs. 4b, 5b, and 6b show, on expanded scales, the portion of the curves near the transformation temperatures. The curves in Figs. 4c, 5c, and 6c are for specimens quenched from the designated temperature (indicated with arrow on the high temperature curve), and measured at 0°C In general, these specimens were held 24 hours at each temperature. For the one curve in Fig. 6c a holding time of only 10 minutes was used during progressive cooling in order to show departure from equilibrium when insufficient time is allowed.

Several features of these curves are noteworthy. First, the resistance changes discontinuously in Figs. 5 and 6 as the temperature passes through the two-phase region. Here, there is a pronounced hysteresis, particularly with the 54 percent alloy, which is typical of other phase transformations. The discontinuities indicate that the resistance of the ordered phase *at temperature* is higher than that of the disordered phase. This is contrary to classical theory and will be discussed in detail later. On measuring at low temperatures, however, the resistance decreases on ordering in the conventional manner as shown by the curves in Figs. 4c, 5c, and 6c. The discontinuity in the low temperature resistance curves provides a more prominent indication of the transformation temperatures when adequate time is employed, e.g., 24 hours or more at each temperature. Three sets of values representing different degrees of hysteresis are listed in Table II in the columns that refer to electrical resistivity measurements.

Microstructural Analyses

Microstructure studies were made using the thin discs, $\frac{1}{4}$ inch in diameter, which had previously been used for hardness tests.[3] The microstructure characteristic of the disordered state consists of equiaxed twinned grains typical of many single-phase face-centered cubic cubic alloys as in the main body of Fig. 7a. The micrograph also reveals evidence of transformation in the unrepresentative surface layer. A subgrain structure was occasionally noticed in disordered specimens and is reported here though it may not be directly connected with the ordering phenomenon. The size of the subgrains was measured to be about 3×10^4A in diameter.

Three different types of microstructures have been observed for fully-ordered alloys: (1) that in which sharply defined striations are apparent as in Fig. 7b; (2) that in which only general mottling of poorly defined features is apparent; and (3) that in which irregular grains have formed at boundaries of the parent grains as in Fig. 7c. All of these were single-phase structures as shown by x-ray diffraction patterns made for the polished surface of the specimens. Consequently, the above features were considered to represent supplementary rather than basic features of the ordering process. The microstructure of the metal in the highly ordered state depends strongly on the method by which the state was reached as was discussed in a previous paper.[3]

Microscopic examinations were also made on specimens that had been heated at various temperatures in

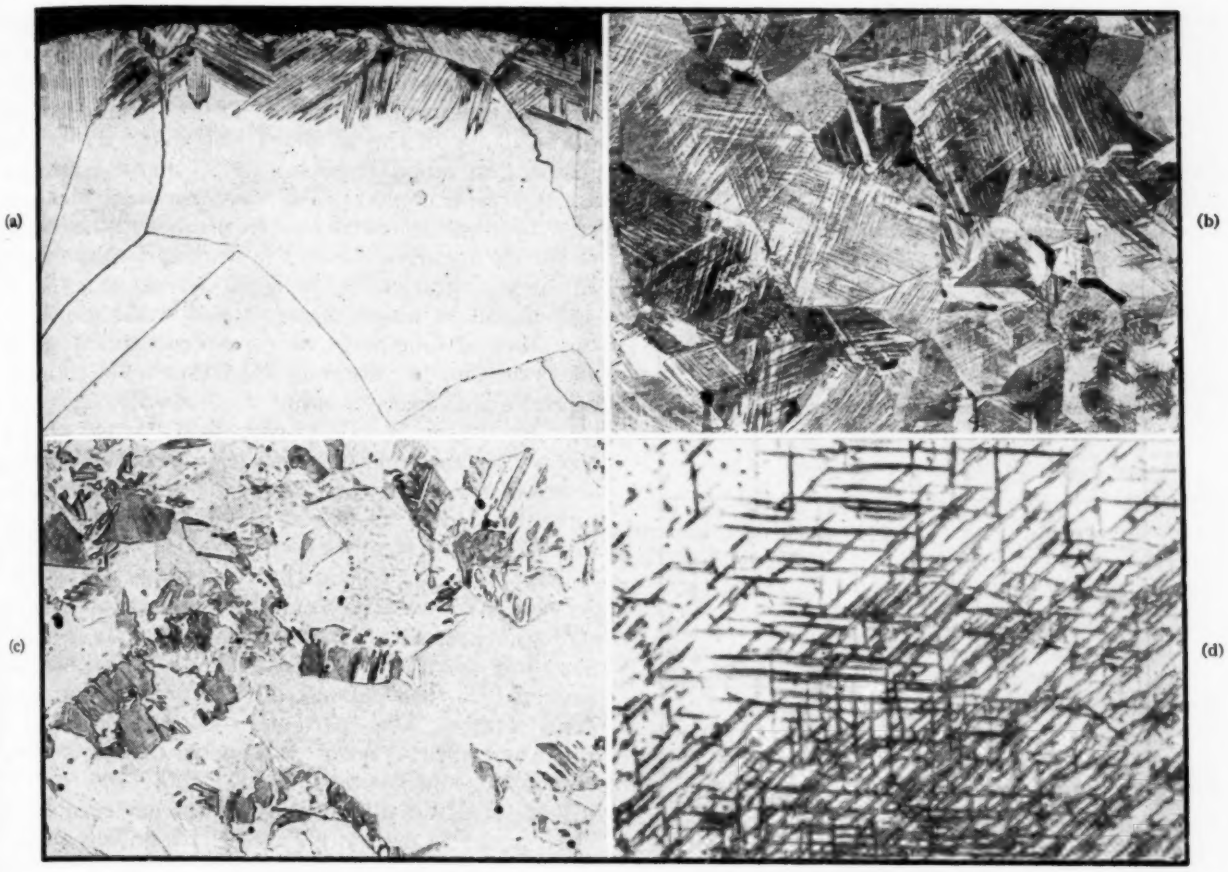


FIG. 7. Microstructures of Co-Pt alloys at equilibrium. Etch: NaCN—aqua regia. (a) Grain structure typical of disordered 48 percent Co alloy. Sample cooled directly from 950°C to 822°C and held 25 hours during which time the unrepresentative surface layer had ordered. 100×. (b) Striations in fully ordered 54 atomic percent Co alloy aged 1000 hours at 600°C. 100×. (c) Irregularly-shaped, recrystallized nodules in ordered 48 atomic percent Co alloy aged 1000 hours at 500°C. 250×. (d) Widmanstätten pattern of ordered phase in disordered matrix of 54 atomic percent Co alloy aged 50 hours at 728°C. 1500×.°

and near the two-phase region. While structures similar to those mentioned above were usually observed, occasionally structures consisting of two discrete phases were detectable. This is shown in Fig. 7d. Here may be seen distinct regions of one phase (presumably ordered), which form a definite Widmanstätten pattern within a matrix phase (presumably disordered). A Debye-Scherrer photograph made on the polished surface of this specimen showed the coexistence of the random face-centered cubic pattern and the ordered tetragonal pattern as in Fig. 2b. This represents metallographic confirmation of the existence of a two-phase region. Stereographic analysis of the Widmanstätten figure shows that the observed traces are due to plate-like particles, which are parallel to {110} matrix planes.

The results of microscopic examination were also used to establish the boundaries of the phase fields. By noting the temperatures at which all of the features associated with ordering are absent, one may establish the location of the boundary between the two-phase region and the disordered region. This boundary appears to be between 823°C and 829°C for the 48 atomic percent cobalt alloy and between 734°C and 739°C for the 54 atomic percent cobalt alloy with the averages listed in Table II.

THE REVISED PHASE DIAGRAM

On the basis of the results reported here, the phase diagram for the Co-Pt system given by Gebhardt and Köster (Fig. 1) can be revised as in Fig. 8. The average values for the several determination techniques in Table II are plotted in the insert of Fig. 8 to give the boundaries of the two-phase field which separates the ordered from the disordered phase. It is assumed that the maximum occurs at 50 atomic weight composition. The rapid descent of the two-phase region on the cobalt-rich side of the ideal composition and the gradual drop on the other are in agreement with the magnetic and resistance measurements of Gebhardt and Köster. Assymetry of the ordering range about the 50 atomic percent composition is not unusual, having been found in several other systems including CuPt, CuPd, and CuAu.[7] Since the existence of a tetragonal disordered phase was disproved, the boundary between "γ_{TD}"

[7] G. Borelius, Z. Elektrochem. 45, 16 (1939).

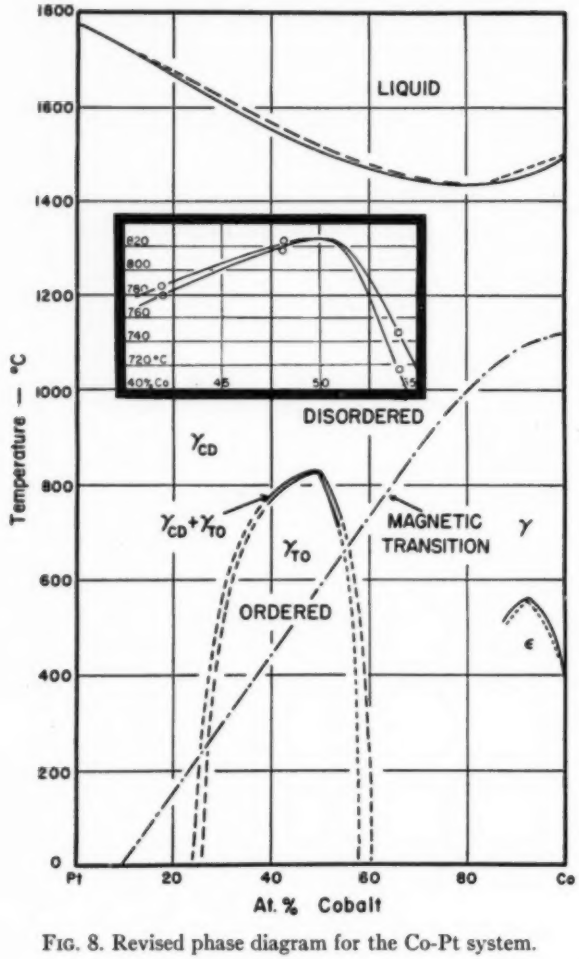


FIG. 8. Revised phase diagram for the Co-Pt system.

and γ_{TO} in Fig. 1, based on inflections in the resistance curves, was omitted. The inflections are attributed to the magnetic Curie temperature rather than to the critical temperatures of order as believed by Gebhardt and Köster. It is well known that the resistivity of a ferromagnetic material below the Curie temperature is much lower than one would expect by extrapolating from higher temperatures. This effect, known since Lord Kelvin's observations in 1851, has been particularly well demonstrated for nickel by Potter and Conybear.[8]

Other revised features of Fig. 8 were based only on the previous work. The phase limits for the transformation in cobalt-rich alloys represent an average of the curves for heating and cooling in Fig. 1 supplemented by an estimated curve (dashed), to complete the fields of two-phase equilibrium between γ and ϵ. Likewise the two phase fields of liquid plus solid γ_{CD} were completed with an estimated curve.

DISCUSSION

The close correlation of the temperatures of the sharp discontinuities in resistance—temperature curves for the alloys between 675–830°C—with those at which the order-disorder transformation range was found by other methods, leaves little doubt that they are associated with the ordering phenomenon. Taylor[9] found a similar discontinuity in Cu_3Pd alloys just above the critical temperature range. The discontinuity and its hysteresis are typical of normal phase transformations that proceed by nucleation and growth. However, the rise in resistivity apparent also in the work of Gebhardt and Köster is contrary to the behavior of most other order-disorder alloys where the resistance falls abruptly at the onset of long-range order. Apparently in Co-Pt alloys the drop in resistance, which one would normally expect to accompany ordering, is obscured at high temperatures where the ordered phase is characterized by a higher resistance than the disordered phase and can be detected only by making measurements at low temperatures. Because of hysteresis, the true temperatures at which the discontinuities occur are somewhat obscure. We may, however, assume that the average of the temperatures at which discontinuities are found on heating and cooling represents the midpoint of the transformation range and thereby arrive at figures which may be compared to those found by other methods as in Table II.

The agreement is fairly good among the various test methods for alloys with 42 percent and 48 percent Co in Table II. X-ray diffraction results on unetched specimens of the 54 percent Co alloy, however, depart appreciably from the results of other techniques. A possible explanation is based on the presence of impurities in the surface layer of the wire specimens, for numerous examples have been found in which small amounts of a third element have markedly shifted the ordering transformation temperatures.[10–12]

The results of the present investigation have more far reached significance than merely providing a revision of the Co-Pt phase diagram. The Debye-Scherrer pattern of Fig. 2b furnishes a number of interesting facts concerning the nature of the ordering reaction. They are here briefly discussed.

1. By proper selection of composition and annealing temperature, it is possible to produce a two-phase structure in which discrete regions of ordered tetragonal lattice and of disordered cubic lattice coexist in equilibrium within individual grains. The equilibrium two-phase pattern shown in Fig. 2b could not be caused by grains which are structurally homogeneous throughout. It is rather the result of an aggregate of grains, each of which is at a state of equilibrium within itself and is heterogeneous with respect to lattice structure. A study of the {200} and {311} lines is most instructive with

[8] H. H. Potter, Proc. Phys. Soc. (London) **49**, 671 (1937).

[9] R. Taylor, J. Inst. Metals **54**, 255 (1934).

[10] A. Komar and N. Volkenstein, Inst. Chem. Gen. (USSR) **16**, 105 (1943).

[11] J. L. Haughton and R. J. M. Payne, J. Inst. Metels **46**, 457 (1931).

[12] R. Hultgren and L. Tarnopol, Trans. Am. Inst. Min. Met. Engr. **133**, 228 (1939).

regards to this conclusion. Whenever a particularly heavy spot occurs on the cubic {200} or {311} line it is matched horizontally by another heavy spot on one or both of the tetragonal {200} or {311} lines. This implies that both cubic and tetragonal lattice structures exist in discrete regions within the individual crystals which make up the wire specimen. The sharpness of the main lattice lines and of the superlattice lines indicates, furthermore, that these discrete regions of order and of disorder were unstrained and large enough to afford sharp Bragg diffraction, i.e., much greater than 1000A in minimum dimension.

2. Within a given crystal, the three orthogonal axes of the tetragonal ordered lattice lie in the directions of the three cubic axes of the parent disordered matrix.† The evidence for this is also given by the horizontal matching of the heavy diffraction spots on the {200} and {311} lines. This effect could not occur with such regularity unless the stated orientation relation obtained in many crystals. Although Fig. 2b gives no information concerning the habit plane of ordered regions forming within a disordered matrix, the tendency to form a Widmanstätten figure, as shown in Fig. 7d, is an expected consequence of the existence of a definite orientation relationship between the lattices. A similar condition was shown to exist in the CuAu superlattice by Dehlinger and Graf.[13]

It should be pointed out that the situation described above is not consistent with a popular concept of order-disorder transformations. To quote a proposed definition:[14] "An order-disorder transformation is a process whereby one structure changes to another without the presence of sharp boundaries in the system at any time." One can hardly say that the transformation involved here is any but that of order-disorder. Yet our evidence proves that sharp boundaries can exist between discrete regions of ordered and of disordered materials both during the process of ordering and in equilibrium. Therefore, a revision of the above concept of the ordering phenomenon and of the mechanism of ordering is necessary. A more detailed analysis of the abundant literature on order-disorder is being prepared for future publication. It is pointed out here that a heterogeneous character of ordering processes has long been assumed by many scientists, but there is copious evidence in its favor.[6,15] The thought that ordering is a homogeneous transformation appears to be a result of such causes as low diffraction rates (when ordering temperatures are low), high rates of nucleation, lack of equilibrium or inadequate experimental techniques.

Table III. Potential strain for various conjugate planes in the disordered and ordered structures.

Plane	Area $(A)^2$	Possible coherency planes	Difference in areas	Percent difference in area
$\{100\}_{Dis.}$	14.10	$\{100\}_D//(100)_O$	−0.2*	−1.42
$\{110\}_{Dis.}$	19.90	$\{100\}_D//(001)_O$	+0.3	+2.13
$\{111\}_{Dis.}$	12.18	$\{111\}_D//(111)_O$	+0.12	+0.99
$(100)_{Ord.}$	13.90	$\{110\}_D//(110)_O$	−0.24	−1.2
$(001)_{Ord.}$	14.40	$\{110\}_D//(101)_O$	+0.10	+0.52
$(110)_{Ord.}$	19.66			
$(101)_{Ord.}$	20.00			
$(111)_{Ord.}$	12.30			

* Sign is based on: $Area_{Ord.} - Area_{Dis.}$

Nature's choice of the {110} matrix planes as the habit plane for the platelike particles of the ordered phase in the disordered matrix satisfies the requirement of minimum lattice strain which has been observed to obtain in many solid state transformations. Since the three orthogonal axes of the crystal lattices of the ordered and disordered phases have been found to be parallel respectively, the areas of conjugate crystallographic planes can be compared in order to find the combination in which the two lattices can be coherent or continuous with minimum strain. Results in Table III show that minimum strain is involved if a $\{110\}_D$ type plane is chosen as the habit plane with the $(101)_O$ plane as the conjugate plane in the particles of the ordered phase.

Wherever two different phases exist together in equilibrium, it is required on the basis of free energy equality that their composition be different from one another. The familiar principle of the tie-line extending across a two-phase region in a binary constitution diagram is a manifestation of this thermodynamic requirement. Thus, ordered and disordered phases coexisting in equilibrium should have slightly different compositions, a condition which has been applied in Shockley's treatment of superlattice theory[16] and evidence for the effect has been reported by Hultgren and Tarnopol,[12] and by Esch and Schneider.[15] Some evidence suggested such segregation in the Co-Pt alloys but the results of the experiments were not sufficiently conclusive.

While the relative amounts of ordered and disordered phases at equilibrium can be determined by the tie-line principle, the degree of order in the ordered phase as determined by a balance between the ordering forces and thermal vibrations of the atoms remains a function of temperature as proposed by classical theory of the superlattice reaction. The data for electrical resistance measured at 0°C as a function of holding temperature (Figs. 4c, 5c, and 6c) clearly show that the degree of order gradually increases (resistivity decreases) with decreasing temperature of equilibrating the specimen

† The axes of the two lattices are approximately parallel. More precisely it will be shown from Table III that the following orientation relationship would be expected to obtain:

$$(110)_D//(101)_O \quad \text{and} \quad [001]_D//[010]_O.$$

This requires a slight tile of 27′ to 43′ depending on c/a for $[100]_O$ and $[001]_O$ from the cube axes. Since only the approximate relationship was determined, subsequent references are made to it alone.

[13] U. Dehlinger and L. Graf, Z. Physik **64**, 359 (1930).

[14] D. Harker, Trans. Am. Soc. Metals **32**, 210 (1944).

[15] U. Esch and A. Schneider, Z. Elektrochem. **50**, 268–274 (1944).

[16] W. Shockley, J. Chem. Phys. **6**, 130 (1938).

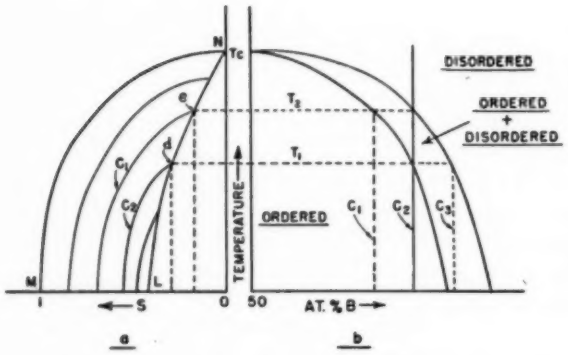


FIG. 9. Scheme for specifying degree of long-range order in two-phase alloys.

after the sudden evolution of order in the transformation range. The diagram in Fig. 9 is proposed as a means for specifying the degree of long-range order of the ordered phase in the range where both disordered and ordered phases coexist in equilibrium. A part of the phase diagram, on the right side of the ideal AB composition, is drawn at the right as Fig. 9b. To the left of this and on the same temperature scale is drawn the curve MN showing the degree of long range order (S) as a function of temperature for the ideal AB composition. Other similar lines are drawn below MN (showing the variation of S with temperature for other compositions), out to that corresponding to LN, beyond which it is not possible to have an ordered phase in the given alloy according to the phase diagram. The line LN is in effect a plot of the degree of order for compositions on the ord.→ord.+dis. boundary as a function of temperature.‡ Each of the lines out to its intersection with LN represents the change of S for the ordered phase with temperature at a given composition.

SUMMARY AND CONCLUSIONS

A study of the equilibrium states of Co-Pt alloys that are susceptible to an order-disorder reaction does not confirm the presence of a previously reported high temperature disordered tetragonal phase and a low temperature body centered cubic phase. The structure of the ordered phase is face-centered tetragonal with an atomic distribution analogous to that of CuAu. This lattice forms from the disordered face-centered cubic phase in such an orientation that the ⟨100⟩ directions are parallel to the ⟨100⟩ directions of the disordered phase. At compositions other than the ideal 50 atomic percent, the ordered phase can exist in equilibrium at high temperatures with the disordered phase. A Widmanstätten pattern of platelike particles of the ordered phase parallel to {110} planes of the disordered matrix is observed when the disordered alloy is held at these temperatures. The temperature range of the two-phase region has been determined for three alloys by x-ray diffraction analyses, microscopic examination, and electrical resistivity measurements. The results for Co-Pt imply that a two-phase region between fields of order and of disorder should be included in equilibrium diagrams of metallic systems. A schematic method has been proposed for describing the degree of long-range order of the ordered phase within the two-phase region.

‡ The curves LN and MN for Co-Pt alloys would be much flatter on top consistent with the high degree of order at the high temperatures shown by x-ray and electrical resistance data.

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

Boride Cathodes*

J. M. LAFFERTY
General Electric Research Laboratory, Schenectady, New York
(Received August 28, 1950)

The thermionic emission properties of the borides of the alkaline-earth and rare-earth metals and thorium have been investigated. These compounds all have the same formula MB_6 and the same crystal structure consisting of a three-dimensional boron framework in whose interlattice spaces the metal atoms are embedded. The valence electrons of the metal atoms are not accepted by the B_6 complex, thus giving rise to the presence of free electrons which impart a metallic character to these compounds. This, together with the strong bonds between the boron atoms in the framework, produces a series of compounds which have high electrical conductivities and high thermal and chemical stabilities—ideal properties for a cathode material. When this structure is heated to a sufficiently high temperature, the metal atoms at the surface evaporate away. They are, however, immediately replaced by diffusion of metal atoms from the underlying cells. The boron frame work does not evaporate but remains intact. This process gives a mechanism for constantly maintaining an active cathode surface. Thermionic emission measurements made on these materials show the rare-earth metal borides to be superior to the others. The highest emission was obtained from lanthanum boride. Its emission constants for the Dushman equation were $\phi=2.66$ volts and $A=29$ amps/cm²/degK². This is higher than the emission normally obtained from thoria. Lanthanum boride has a relatively low evaporation rate corresponding to a latent heat of evaporation of 169 kilocalories per mole. If the hexaborides are operated at high temperature in contact with the refractory metals, boron diffuses into their metal lattices forming interstitial boron alloys with them. When this occurs, the boron framework which holds the alkaline-earth or rare-earth metal atoms collapses, permitting the latter to evaporate. However, the hexaboride cathodes may be operated at high temperatures in contact with tantalum carbide or graphite. Lanthanum boride cathodes are especially useful in applications where high current densities are required. They are also suitable for high voltage applications because they stand up well under positive ion bombardment. Since they are atmospherically stable and activate easily, they have found wide use in experimental demountable systems.

I. INTRODUCTION

IN keeping pace with the trend towards higher and higher frequencies, the tube designer is continually confronted with the problem of extracting larger current densities from cathode surfaces. The problem has become so critical that electron emission efficiency is sacrificed in order to use cathode materials which will supply the requisite current densities and have a reasonable life. In many cases, even the latter is difficult to obtain. Three types of cathodes are now in common use. They are (1) clean metals, (2) metals with contaminated surfaces, and (3) nonmetallic electron emitters.

Type (3) make the most efficient cathodes, but cannot be operated at high direct current density levels because of the excessive evaporation of the materials which contribute to their high thermionic activity.

With regard to type (1), it is found that those elements provided by nature which are the best electron emitters also evaporate most readily. When such metals as cesium, barium, cerium, or thorium are heated to temperatures high enough to give electron emission currents of a few amperes per square centimeter, the evaporation rate becomes so high as to make the cathode useless in high vacuum devices. The refractory metals have poor emission efficiencies, but have more favorable evaporation rates. Figure 12 shows the evaporation rates of various refractory metals when they are heated to temperatures necessary to produce a given current density. The more efficient emitters would be located many decades above the refractory metals on this graph.

If these same metals, cesium, barium, cerium, or thorium are deposited on a refractory base metal such as tungsten in the form of films only one atom thick, these atoms become polarized to some extent and reduce the work function of the emitter below the value characteristic of the adsorbed metal in bulk. It is found that the evaporation rate of the atoms from the monolayer is much less than from the metal in bulk. These emitters may be classified as type (2) mentioned above. They form a much more desirable cathode for high current density applications.

Long life may be obtained from these cathodes providing some mechanism is provided for replacing those atoms in the monolayer which evaporate away or are sputtered off by high velocity positive ion bombardment due to residual gas in the tube. Considerable thought and ingenuity has been devoted to methods of accomplishing this. Some methods use an independent source which continually evaporates the active material onto the cathode surface. Other methods depend on the continuous decomposition of a compound to supply the active material. The diffusion rates of active materials through various substances have also been used in maintaining the monolayer. Still other methods use various combinations of these.

All of these methods have one inherent disadvantage in that there is no means for automatically regulating the rate of replacement of the active material. To insure an active cathode at all times, the rate of replacement must exceed the rate of loss from the monolayer. This results in a waste of active material.

It has been found possible to overcome this difficulty and to provide a cathode with many other advantageous properties by embedding the alkaline-earth or rare-

*This paper was presented at the Autumn Meeting of the National Academy of Sciences in Schenectady on October 11, 1950.

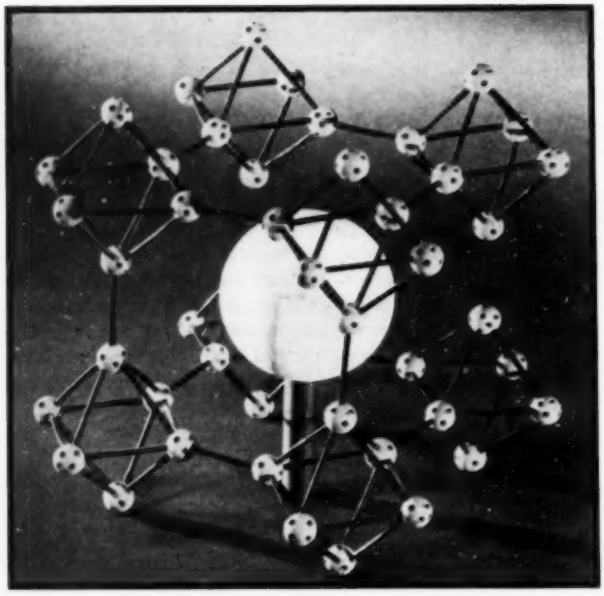


FIG. 1. The hexaboride crystal structure. The boron atoms form a three-dimensional framework which surround the large metal atoms.

earth metals or thorium in a matrix of boron. An understanding of how this is accomplished may be easily obtained by considering the unusual properties of a series of interstitial compounds of boron and the alkaline-earth or rare-earth metals or thorium. These all form borides with the same formula and have the same crystal structure. Although some of these compounds were made over 50 years ago and their crystal structure determined nearly 20 years ago, they appear to have found little application other than to arouse the curiosity of the crystallographer.

II. HEXABORIDE PROPERTIES

The alkaline-earth metals, the rare-earth metals, and thorium form borides of the type MB_6. These compounds all have the same cubic crystal structure as shown in Fig. 1. The small boron atoms form a three-dimensional framework structure which surrounds the large metal atoms. The boron framework is made up of octahedra, one at each corner of the cube. These are bonded together at their apexes. Each boron atom has four adjacent neighbors in its own octahedron and another neighbor in the direction of one of the cubical main axes, thus giving a homopolar lattice structure with the coordination number 5. Each boron atom has its three valence electrons distributed over five bonds.

The metal atoms trapped in the boron cages have a coordination number of 24. It would appear, however, that there are no valence bonds between the metal atoms and the surrounding boron atoms. The valence electrons of the metal atoms thus become free electrons and impart a metallic character to the compounds.

It would appear from Table I that the lattice constants of the hexaborides are determined primarily by the boron frame work structure and not by the size of the embedded metal atoms. The radii of the metals, which can be calculated from these lattice constants, vary but little, notwithstanding the fact that the radii of these atoms in the elementary as well as in the ionized state have very different sizes. It would thus appear that the lattice constant of the three-dimensional framework remains substantially constant during the isomorphous exchange of the metal atoms as long as they remain smaller than the interlattice spaces. If the metal atoms become larger, the framework will expand to accommodate them as is shown by the BaB_6.

The strong binding forces between the boron atoms lead to a series of compounds which are very refractory, with melting points above 2100°C. The melting points of the hexaborides are characteristic of the boron structure and not of the metal atoms. Table I shows that the variation in the melting points is less than 4 percent.

The borides are very stable chemically; moisture, oxygen, and even hydrochloric acid do not react with them. The hexaborides are opaque. The alkaline-earth metal borides are black and brownish black. The rare-earth borides are various shades of blue and purple. When dry, LaB_6 is a reddish purple, but turns a deep red when moist.

TABLE I. Lattice constants[a] and melting points of the hexaborides.

Boride	a_0	mp
CaB_6	4.145A	2235°C
SrB_6	4.19A	2235°C
BaB_6	4.28A	2270°C
LaB_6	4.145A	2210°C
CeB_6	4.129A	2190°C
ThB_6	4.15A	2195°C

[a] M. V. Stackelberg and F. Neumann, Z. physik. Chem. **19B**, 314 (1932).

The metallic character of these compounds is evident from their high electrical conductivity. Figure 2 shows the specific resistance of well-sintered LaB_6 as a func-

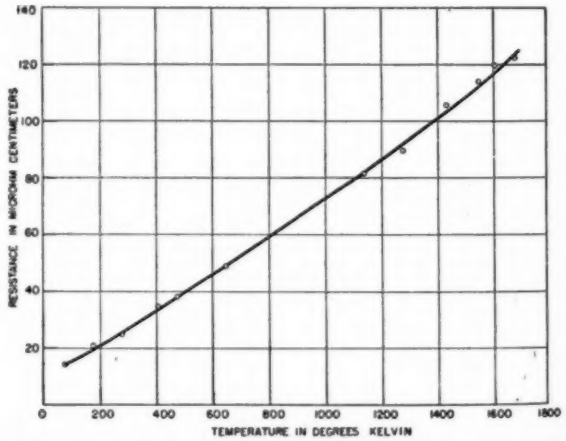


FIG. 2. The specific resistance of sintered LaB_6 as a function of the absolute temperature. The sample was sintered at 1850°C for ½ hour.

tion of temperature. The resistance is approximately the same as that of lead. The resistance-temperature coefficient is positive, like that of a metal. This is true of all the hexaborides. Calcium boride has the highest specific resistance of all the hexaborides, as might be expected because of the small diameter of the calcium atom.

The thermoelectric power of LaB_6 measured against annealed Baker's platinum was found to be 8 microvolts per degree centigrade. These measurements were made over the temperature range 100 to 300 degrees centigrade. The polarity was such that the electron current flow was from the cold junction to the hot junction through the boride.

Hall effect measurements[1] made on LaB_6 also show its metallic properties. A LaB_6 bar 1.261 by 0.588 by 0.151 centimeters was made from pressed powder and sintered in vacuum at 1800°C for 15 minutes. The density was 2.61 grams/cm³. The specific resistance was found to be $\rho = 57 \times 10^{-6}$ ohm centimeters and the Hall coefficient $R = 7.7 \times 10^{-12}$ volt centimeters/ampere gauss. The mobility of the current carriers is calculated from the relation $(R/\rho) \times 10^8$ and found to be 13.5 centimeters/second per volt/centimeter. This value is not uncommon for metals. The density of the carriers is given by the relation $n = 6.3 \times 10^{10}/R$ and found to be 8.2×10^{21} per cubic centimeter or 6.4×10^{23} per mole of LaB_6. This is approximately Avogadro's number, indicating that there is one free electron for every lanthanum atom. In their magnetic investigations of LaB_6, Klemm, Schüth, and Stackelberg[2] found triply ionized lanthanum.

When the hexaborides are heated to a sufficiently high temperature, the metal atoms at the surface evaporate away. They are, however, immediately replaced by diffusion of metal atoms from the cells below. The boron framework does note vaporate, but remains intact. It is found that this diffsion of metal atoms to the surface occurs only when there is a vacancy at the surface. This property keeps evaporation losses at a minimum and at the same time provides a mechanism for constantly maintaining an active cathode surface. This, together with the high electrical conductivity and high thermal and chemical stability, gives ideal properties for a cathode material.

III. CHEMICAL PREPARATION

The borides of Ca, Sr, and Ba were first obtained by Moissan and Williams[3] by reducing the borates with aluminum in an arc furnace. An improved method is given by Andrieux.[4] He electrolyzed a molten mixture of the borate and fluoride of the metals in question at 1000°C. Stackelberg and Neumann,[5] in their study of the MB_6 crystal structure, have also prepared the borides of the alkaline earth and rare-earth metals by the method of Andrieux.

In our work, the borides were prepared from the metal or the metal hydride and amorphous boron. Fresh metal, in the form of powder or filings, was mixed with the amorphous boron powder and pressed into bars 1 by 1 by 6 inches with a 50-ton press. These bars were placed in a graphite cruicible and packed with boron powder. This was fired in an electric furnace in a pure dry hydrogen atmosphere for 1 hour at 1375°C. The boron powder was removed, and hydrogen firing was continued in a graphite crucible in an induction furnace for 20 minutes at 1800°C. The boride thus obtained was crushed, washed in HCl, rinsed in water, and ground dry in a steel ball mill. The powder thus obtained was washed in concentrated hydrochloric acid to remove any free metal, metal oxide, and boric oxide. It was then washed in distilled water several times until all traces of the acid were removed. Small amounts of free boron could be removed to a large extent by stirring the powder in distilled water and then decanting off the water before the powder completely settled out. (The rare-earth borides are approximately 2.5 times as dense as amorphous boron.)

The powder is then dried and is ready for use. It may be pressed into any desired shape and sintered in hydrogen or vacuum. Sintering at 1375°C for 15 minutes gives a material which is soft enough so that it may be machined. Sintering at 1800°C for 15 minutes gives a material which is very hard and about 0.5 the theoretical density.

The metals used in preparation of the borides were quite pure, usually about 99 percent with traces of Fe, Cr, Ni, and Si. The boron, however, was not very pure and contained large amounts of Mg, Si, and Al with traces of Pb, Sn, Fe, Mn, and Cu. A large amount of these impurities are distilled out in the preparation of the boride; and spectrographic tests made on the final boride show that silicon is present, but only traces of Fe, Al, Mg, and Mn.

IV. ELECTRON EMISSION CONSTANTS

In measuring the electron emission constants of the borides, one is confronted with two problems. One is the establishment of temperature scales for the various materials and the other is the prevention of their coming in contact with metals which react chemically with them (Section V) when they are heated. These difficulties were avoided by making a boride cathode in the form of a self-supporting hollow rod, as shown in Fig. 3. The three tungsten coils which surround the rod are used for heating it and for collecting the electron emission from it. The two end coils are connected in series. The heater currents through them and the center coil

[1] The writer is indebted to Dr. W. C. Dunlap, Jr., of this laboratory for these data.

[2] Klemm, Schüth, and Stackelberg, Z. physik Chem. **19B**, 321 (1932).

[3] H. Moissan and P. Williams, Compt. rend. **125**, 629 (1897).

[4] L. Andrieux, Compt. rend. **182**, 126 (1926); **184**, 91 (1927); **186**, 1537 (1928); **186**, 1736 (1928).

[5] M. v. Stackelberg and F. Neumann, Z. physik Chem. **19B**, 314 (1932).

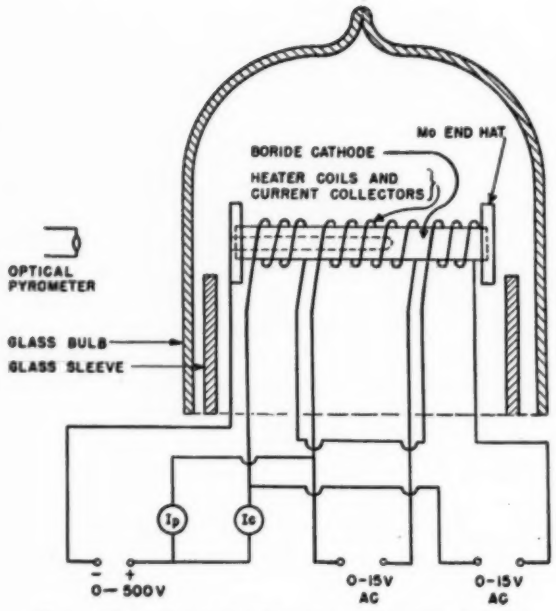


FIG. 3. Experimental tube setup for determining the emission constants of sintered borides.

are adjusted so that the rod, as viewed through the hole, appears to be uniformly bright throughout the major portion of its length. Since the radiation from the cavity is sensibly blackbody radiation, the temperature may be read directly with a pyrometer without emissivity correction. Further, since the rod is heated by radiation from the surrounding coils, its temperature is uniform throughout and the temperature as measured at the inside cavity walls is the same as that at the outer surface.

The three coils are made positive with respect to the rod and are used to collect the electron emission current. The electron current collected by the center coil from the uniform temperature portion of the cathode is measured by the I_p meter. The end coils are connected to the positive voltage source through a similar meter I_c and serve as guard rings, collecting the emission current from the end portions of the rod. The construction and processing of the tubes used in the present measurements follows.

The boride rod was 180 mils in diameter and $1\frac{1}{8}$ inches long. The rod was made by sintering the pressed boride powder for 15 minutes at 1375°C. The surface was then ground in a centerless grinder to give a uniform diameter of 180 mils. A 93-mil hole was drilled in one end to a depth of $\frac{3}{4}$ inch. The rod was cut to a length of $1\frac{1}{8}$ inches and placed in position between the molybdenum end hats. The heater coils were made from 20-mil tungsten wire wound on a 307-mil diameter mandrel 28 turns per inch. The center coil was $\frac{1}{2}$ inch long and the end coils were $\frac{1}{4}$ inch long.

The tube was baked out for one hour at 450°C and the cathode outgassed on the pump by operation at 1500°C to 1600°C. During these operations, a glass sleeve (Fig. 3) protected the bulb from any discoloration at the place through which pyrometer measurements were to be made. Two barium getters were flashed and the tube sealed off. The tube was operated several hours at a temperature greater than that at which data was taken, to insure stable emission.

The electron emission from the various borides was measured as a function of their temperature with a constant potential of 500 volts between cathode and collector. Richardson plots were made from these data, using the thermionic emission equation

$$\log_{10} I/T^2 = \log_{10} A - 11600\phi/2.303T, \qquad (1)$$

where I is the current density in amperes per cm², T the absolute temperature in degrees Kelvin, A is an empirical constant expressed in amperes per cm² per degree Kelvin², and ϕ is the conventional work function expressed in volts.

TABLE II. Hexaboride emission constants as determined from the data in Figs. 4 and 5.

Boride	A(amps/cm²/degK²)	ϕ (volts)
CaB_6	2.6	2.86
SrB_6	0.14	2.67
BaB_6	16	3.45
LaB_6	29	2.66
CeB_6	3.6	2.59
Misch metal boride	14	2.64
ThB_6	0.5	2.92

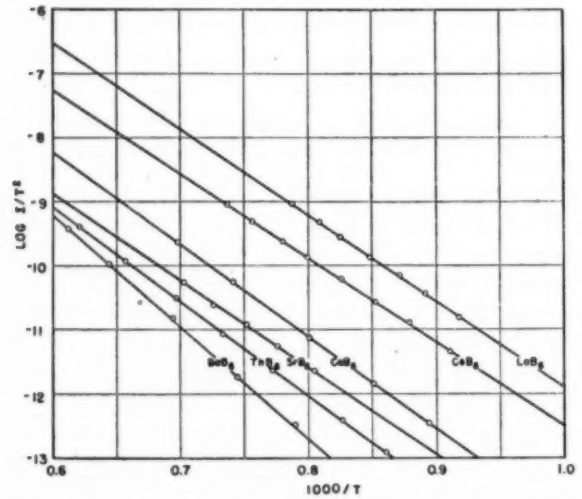


FIG. 4. Richardson plots for the alkaline-earth, rare-earth, and thorium borides. The emission constants are given in Table II.

Figure 4 shows a series of Richardson plots for the alkaline-earth, rare-earth, and thorium borides. Straight lines were obtained in all cases, showing ϕ to be substantially independent of temperature. It will be noted that the rare-earth borides are better emitters than the alkaline-earth or thorium borides. Values of A and ϕ obtained from Fig. 2 are given in Table II.

Figure 5 shows the emission obtained for some mixed borides of the rare-earth metals. X-ray analysis of the crystal structure of these mixed borides shows that they form solid solutions over the entire composition range. Misch metal, which is an alloy of all the rare-earth

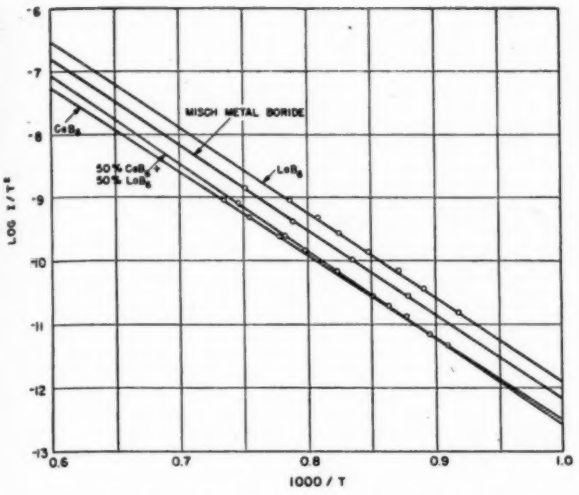


FIG. 5. Richardson plots for some mixed borides of the rare-earth metals. Misch metal boride is a hexaboride made from misch metal—an alloy of the rare-earth metals.

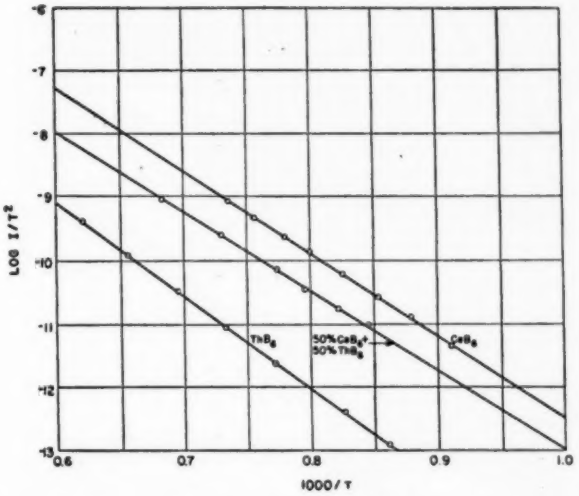


FIG. 7. Richardson plots for mixed borides of cerium and thorium.

metals, forms a solid solution of mixed borides with an electron emission only slightly less than that of LaB_6. This is of commercial interest since misch metal is considerably cheaper than lanthanum metal.

Figure 6 shows the emission obtained from solid solutions of cerium and barium borides of various compositions. Figure 7 shows similar data for cerium and thorium borides. It will be observed that the mixed borides give emission currents which fall within the limits of emission established by the pure components. This result is in accordance with the theoretical considerations stated in Sec. VII.

It is of interest to compare the emission from LaB_6, the best boride emitter, with some of the conventional cathode materials. This is shown in Fig. 8. The emission from LaB_6 is surpassed only by barium oxide at low temperatures. In the high current density range, LaB_6 is superior to any of the conventional materials now available.

V. CATHODE PREPARATION AND BASE MATERIALS

When the hexaborides are heated in contact with refractory metals such as W, Mo, Pt, Cb, or Ta, the boron atoms diffuse into these metal lattices, taking up positions in the interstices and forming interstitial boron alloys with them. When boron diffusion starts, the boron frame work which holds the alkaline-earth or rare-earth atoms collapses, permitting the latter to evaporate. This process will continue until all of the interstices of the base metal are filled with boron or until all of the boron from the hexaboride coating is used up.

The rate at which this diffusion process proceeds is indicated in the following experiment. A filament was

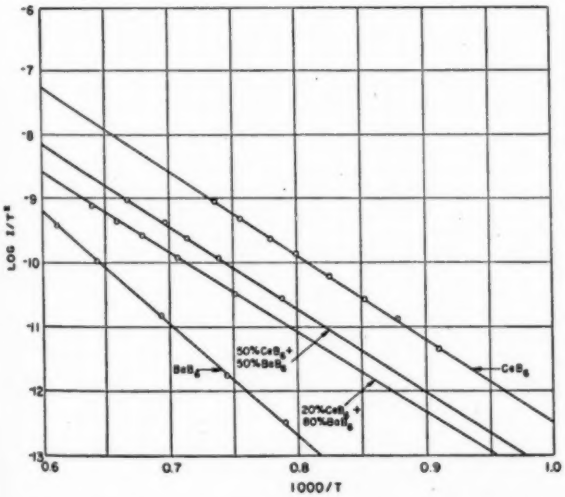


FIG. 6. Richardson plots for mixed borides of cerium and barium in solid solution.

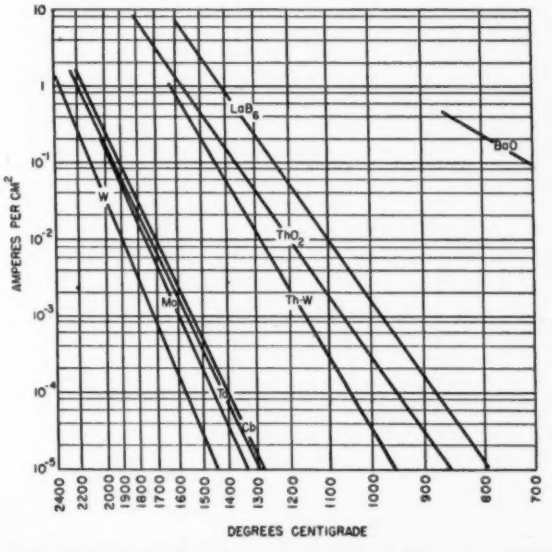


FIG. 8. Electron emission of some common cathode materials and lanthanum boride plotted as a function of the temperature.

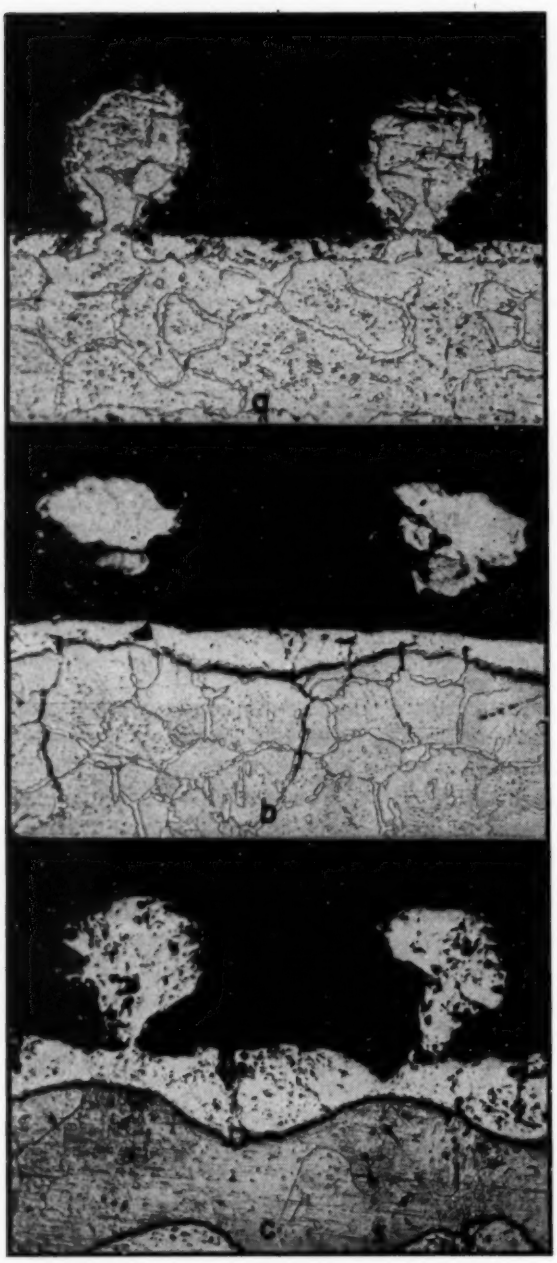


FIG. 9. Photo-micrographs showing cross section of an overwound tantalum filament. In (a) the filament has been carbonized at the surface to a depth of approximately one mil. In (b) the carbonized filament has been heated to 2000°C in contact with boron. In (c) an uncarbonized filament was given the same treatment as in (b).

made with a 20-mil diameter tungsten wire core overwound with 10-mil diameter wire with a 10-mil space between the overwound turns. The space between the turns was filled with LaB_6. The cathode was operated at 1515°C and gave very good emission. At the end of 48 hours, however, the emission had ceased and there was no LaB_6 visible on the overwound tungsten wire. The bulb was covered with a brown deposit which was mostly lanthanum. The tungsten showed a slight swelling and was very brittle, crumbling into very small pieces when handled. A microscopic examination showed many small cracks in the metal. A spectrographic analysis of the filament showed only tungsten and boron present, lanthanum being entirely absent.

Looking at this problem from the interstitial solid solution point of view, an effective way to prevent boron from diffusing into the base metal is to previously fill the interstices with small atoms before it is brought in contact with the boride at high temperatures. The metalloids, boron and carbon, are most suitable for this purpose.

If the cathode is to be made by completely boronizing or carbonizing the base metal, the rate of diffusion of the metalloids in the base metal is of little importance, since the evaporation rates of the metalloids are negligible at the cathode operating temperature. However, completely boronizing or carbonizing the base metal always results in a very brittle alloy; and for some applications, such as filaments, this may prove to be a severe limitation in handling. It would be better, therefore, just to boronize or carbonize the surface of the base metal, leaving a ductile core. In this case, the diffusion rate of the metalloid into the base metal becomes important. If the diffusion is appreciable at the operating temperature of the cathode, the volume near the surface of the base metal will be depleted of metalloid atoms, leaving vacant interstices into which the boron will diffuse from the hexaboride resulting in its dissociation. It follows that boron cannot be used in cases where a ductile core is required because of its high diffusion rate in the base metal. If this were not so, the boride would require no protective barrier.

Tests made on the refractory metals show that the diffusion of boron and carbon is slowest in tantalum. Overwound tantalum filaments were carbonized by packing them in finely powdered sugar charcoal in a graphite crucible and heating in vacuum for one minute at 2300°C. This gives a tantalum carbide coating of about one or two mils thickness.[6] At 1500°C, this coating is extremely stable and prevents the boron from diffusing into the tantalum.

Figure 9 shows some photo-micrographs of a 20-mil diameter tantalum wire overwound with 10-mil diameter tantalum wire. These were cross sectioned, polished, and etched, to show the carbide and boride surfaces. Figure 9a shows a sample which was carbonized at the surface as described above. Figure 9b shows a carbonized sample which was heated for 5 minutes at 2000°C in contact with boron. Figure 9c shows a sample which was not carbonized, but was heated for 5 minutes at 2000°C in contact with boron. It can be seen that in the unprotected sample, the boron has diffused entirely through the overwinding and nearly half-way through the core wire.

[6] The carbonizing could have been done equally well or better in a hydrocarbon atmosphere.

An unprotected overwound tantalum wire coated with LaB_6 and operated at 1515°C for 144 hours showed diffusion of boron into the tantalum for a depth of approximately 3 mils. On carbonized tantalum, no evidence of boron diffusion was found under similar operating conditions.

It was also observed that the borides do not react with graphite. Many cathodes have been made by putting the boride in graphite cups or by sintering it on graphite cylinders and plates which have scored or undercut surfaces. Graphite may also be used as a boron barrier on filaments.[7]

Although the borides will sinter by themselves at temperatures as low as 1400°C, they do not adhere well to metal or ceramic surfaces. For this reason, they cannot be applied to cathode base materials in the conventional way that the carbonates or thoria are applied. The borides may be applied to form flat cathode surfaces by drilling, grooving, or fastening a coarse mesh to the base material. These methods provide surfaces into which the boride may lock. Filamentary boride cathodes are made by overwinding the filament core wire with another wire $\frac{1}{2}$ its diameter. The space between turns of the overwound wire is made equal to its diameter, and is then filled with boride.

The boride powder is mixed with amyl acetate to form a thin paste which is painted on the cathode base material. The cathode is then sintered in vacuum for 15 minutes to one hour at temperatures ranging from 1700°C to 1800°C.

Solid boride cathodes may be made in the form of disks, rods, etc., by pressing the powder into the desired shape and sintering in vacuum. Unusual shapes may be formed by the use of graphite molds. The mold is filled with boride paste and sintered. After firing, the cathode can be easily removed from the mold, since there will be some shrinkage of the boride and there will be no bonding between the latter and the graphite.

VI. EVAPORATION RATE OF LANTHANUM BORIDE

The usual method for determining the evaporation rates of highly refractory metals is to observe the rate of loss in weight of filaments operated at constant temperatures. This method presents several difficulties when applied to the borides. At the present state of the art, these compounds cannot be made into wires. If a refractory metal wire is coated with the boride, there is always danger of chemical reaction between the two at high temperatures. If sintered boride rods are used, there is the problem of uniform heating, and an accurate determination of the temperature is difficult. Another factor which makes the rate of loss in weight method undesirable when applied to cathodes is that all measurements must be made at temperatures considerably higher than the operating temperature. The data obtained must then be extrapolated over several decades to apply at normal operating temperatures. This may introduce considerable error.

[7] Mr. M. L. Perl of the Electronics Department of the General Electric Company has developed a process for depositing a tenacious layer of carbon on ductile tantalum wire.

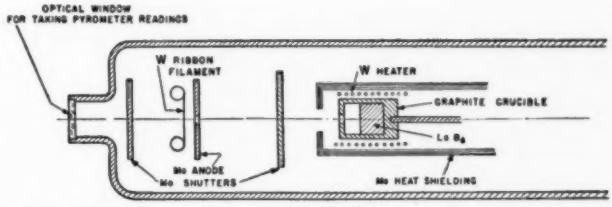


FIG. 10. Schematic diagram of the tube used for measuring the evaporation rate of lanthanum boride.

The method described below has the advantage that the true temperature of the sample may be read directly with an optical pyrometer; no weighing or chemical analysis is necessary, and readings may be taken near the operating temperature in relatively short times. This method may be used to measure the evaporation rate of any material whose evaporation product forms an active thermionic surface.

A schematic diagram in Fig. 10 illustrates the method used. The LaB_6 is placed in a graphite crucible which does not react with it at high temperatures. There is a small opening in the end of the crucible through which the lanthanum can evaporate. The crucible is heated by a closely wound tungsten coil which surrounds it. Several layers of thin molybdenum foil are placed around the whole assembly to act as heat shielding and to give a uniform distribution of temperature throughout the crucible. The lanthanum evaporating from the crucible passes through the hole in the molybdenum anode and condenses on the tungsten ribbon. This ribbon is operated at about 800°C so that the rate of evaporation of lanthanum atoms from a single layer on the surface is negligible. The anode is made 100 volts positive with respect to the ribbon.

When the crucible has been heated for a time sufficient to reach an equilibrium temperature, the molybdenum shutter is opened and the lanthanum allowed to condense on the ribbon. The electron emission from the ribbon to the anode is recorded as a function of time on a photoelectric recorder. The length of time for the emission current from the ribbon to reach a maximum is measured. This corresponds to the time required for a monolayer of lanthanum to deposit on the ribbon. The temperature of the LaB_6 is measured with an optical pyrometer by opening both shutters and reading the temperature inside the crucible. Only a correction for the window need be made, since the radiation from the crucible is essentially blackbody radiation.

The rate of evaporation of lanthanum from the LaB_6 can be computed as follows. Clausing[8] has shown that the number of atoms per second per unit solid angle leaving the crucible in the direction of the hole axis is νr^2, where ν is the number of atoms striking the crucible wall per square centimeter per second and r is

[8] P. Clausing, Z. Physik **66**, 471 (1930).

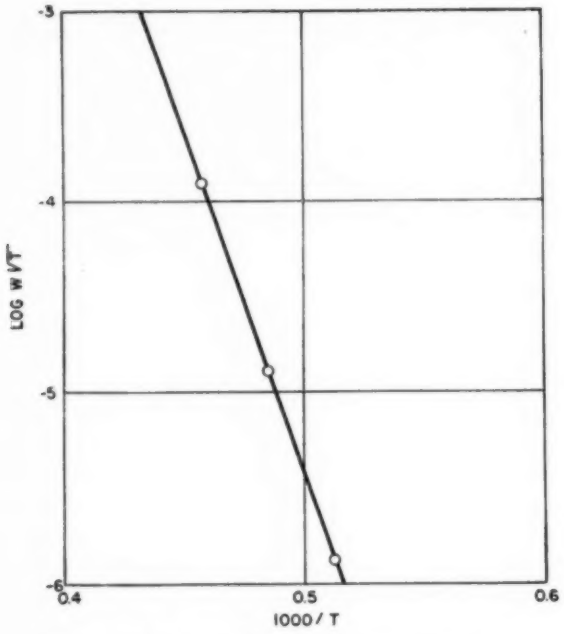


FIG. 11. Evaporation rate data for lanthanum boride.

the radius, in centimeters, of the hole in the crucible. The number of atoms per square centimeter falling on the ribbon in time t is $\nu r^2 t/R^2$, where R is the distance between the ribbon and the opening in the crucible. If N_0 is the number of lanthanum atoms per square centimeter required to give a monolayer on the tungsten ribbon, then the number of atomic layers deposited on the ribbon in time t is $\nu r^2 t/R^2 N_0$. If t is the time required to deposit a monolayer, then on equating this expression to unity and solving for ν, it becomes

$$\nu = R^2 N_0 / r^2 t. \tag{2}$$

This is the number of lanthanum atoms leaving the LaB_6 surface per square centimeter per second. Expressed in grams per square centimeter per second, this quantity becomes

$$W = MR^2 N_0 / N_A r^2 t, \tag{3}$$

where M is the atomic weight of lanthanum and N_A is Avogadro's number.

The uncertainty in N_0 will cause the greatest error in calculating W from Eq. (3). The absorbed lanthanum atoms will no doubt arrange themselves in conformity with the underlying tungsten lattice. However, the lanthanum atoms are too large to form a one-to-one correspondence with the tungsten atoms. A one-to-two or one-to-three relationship would be more probable. Further, it is believed that the maximum electron emission occurs when the concentration of the absorbed atoms is somewhat less than a complete monolayer. Also, the actual surface area of thermally etched tungsten will be greater than the projected area because the surface consists of many crystal faces inclined to the original surface. Since the interest here is in determining lanthanum boride cathode life, it will be better to assume a value for N_0 which is somewhat large. If the closest possible packing of the lanthanum atoms on a plane surface is assumed, a value of 8.33×10^{14} atoms per square centimeter is obtained. (This is not appreciably different than 7.13×10^{14}, which is the number of absorbed atoms per square centimeter for a one-to-two correspondence on dodecahedral tungsten crystal faces.) If this value is multiplied by 1.3 to allow for orientation of the tungsten crystal faces,[9] a value of 1.08×10^{15} atoms per square centimeter is obtained for N_0.

TABLE III. Evaporation rate data for lanthanum boride.[a]

T (°K)	t (sec)	W (gm/cm²/sec)
2183	2700	2.66×10^{-6}
2063	25,800	2.78×10^{-7}
1953	248,400	2.89×10^{-8}

[a] $R=6.47$ cm. $r=0.0381$ cm.

Table III shows the time required to obtain maximum emission at various crucible temperatures and the evaporation rates computed from Eq. (3).

The evaporation rate of most substances may be expressed by the relation

$$\log W = C - \tfrac{1}{2}\log T - (B/T), \tag{4}$$

where W is the rate of evaporation in grams per square centimeter per second at a temperature T degrees absolute, and C and B are empirical constants. If the data in Table III are plotted as $\log W(T)^{\frac{1}{2}}$ *vs* $1/T$, a straight line should result if Eq. (4) applies. This is shown to be the case in Fig. 11. The following values of C and B were obtained from this curve:

$$C=13 \quad \text{and} \quad B=36{,}850°\text{K}.$$

It should be noted that the value of B, which corresponds to a latent heat of evaporation of 169 kilocalories per mole, is not dependent on the geometry of the setup or the value assumed for N_0, but depends only on T and t.

A comparison of the evaporation rate of LaB_6 with the refractory metals when used as electron emitters is shown in Fig. 12. These curves shown the evaporation rates at temperatures which will give the emission current densities shown by the abscissas.

VII. ACTIVATION AND EMISSION PHENOMENA

Boride cathodes require no activation in the usual sense of the word. If a LaB_6 cathode, for example, is heated to 1500°C or 1600°C for a few minutes for outgassing, it is found to be completely active. In fact, it is impossible to outgas the cathode properly without making it active. However, it was found that LaB_6 cath-

[9] L. Tonks, Phys. Rev. **38**, 1030 (1931), has calculated a value of 1.225 for the ratio of the exposed crystal face area to the projected area for dodecahedral faces assuming random crystal orientation and minimum face excess.

odes, which had been washed in HCl and not previously heated in vacuum, were inactive at 1100°C.

The activation of LaB_6 was investigated further in a special tube shown in Fig. 13. LaB_6 powder was placed in a graphite cup heated by radiation from a surrounding tungsten coil. A graphite anode was placed inside the cup near the LaB_6 as shown. The anode contained a small hole through which the temperature of the LaB_6 could be measured with an optical pyrometer.

The LaB_6 was heated at an activation temperature for a period of time, and its thermionic condition was checked frequently by dropping to a testing temperature of 1115°C and measuring the saturated emission. The saturated emission at 1115°C is plotted against the activation time for activation temperatures of 1375°C and 1430°C. The activation rate is approximately three times as large at the higher temperature.

An active LaB_6 cathode opened up to the air and then evacuated again will be inactive when first heated to 1115°C. If it is heated for a few moments at 1500°C or 1600°C, it will return to its original activity.

When LaB_6 is operated in vacuum for a long period of time at a high temperature, a brown deposit collects on the glass walls of the tube. An x-ray analysis shows this material to be amorphous. It is soluble in hydrochloric acid and a spectrographic analysis shows it to be mostly lanthanum. A microscopic examination of the LaB_6 surface shows it to be clean with no residue present.

From these facts, it is concluded that the thermionic activity of the cathode is caused by the presence of lanthanum at the surface of the LaB_6. Whether this is a layer of lanthanum atoms on a LaB_6 surface or simply a LaB_6 surface is somewhat problematical because of the high concentration of lanthanum atoms in LaB_6. If LaB_6 dissociated in the usual manner on heating, a residue of boron would be left on its surface, since the vapor pressure of boron is extremely small compared with lanthanum. Since none is found, it is concluded that lanthanum diffuses out through the boron framework and evaporates, while the boron structure remains intact. After prolonged operation at high temperatures, the LaB_6 will become so depleted in lanthanum that the boron framework will collapse, leaving a boron coating on the cathode.

The nature of the thermionic activity and the lanthanum atoms at the surface of the LaB_6 was investigated further by operating the cathode in mercury vapor. A graphite cylinder coated with LaB_6 was heated by radiation from a tungsten coil within. The anode was connected to the positive side of a 110-volt line through a variable resistance.

With the cathode operated at 1500°C, the anode current would increase with a decrease in the anode series resistance. The anode drop was approximately 10 to 15 volts. When the anode current approached the temperature limited emission of the cathode, the anode voltage would rise and, at about 20 to 30 volts, the anode current would drop rapidly to a very low value with the anode voltage rising to nearly 110 volts. If the anode series resistance was now increased again to its former value, the cathode would activate immediately and the anode voltage would drop to its former value. This procedure could be repeated many times, always with the same results. When the cathode temperature was reduced to 1200°C, similar results were observed except that the cathode would not reactivate when the anode series resistance was increased again. The cathode could be activated again by heating momentarily to 1500°C.

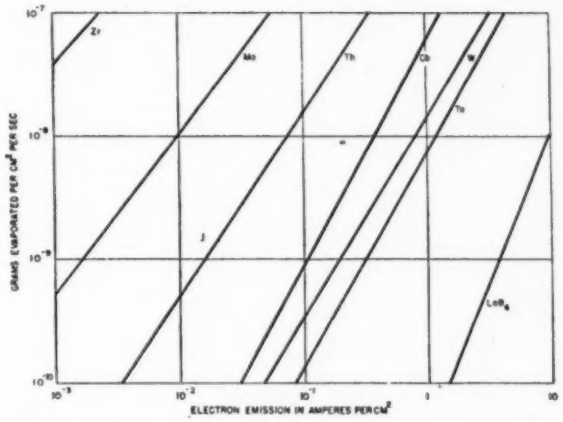


FIG. 12. Evaporation rate of various materials when they are heated to a temperature which will give the electron emission shown.

From this, it is concluded that when the mercury ions exceed energies of approximately 20 electron volts, they readily sputter off the lanthanum atoms from the surface of the LaB_6 and make the cathode inactive. However, at 1500°C, the lanthanum atoms diffuse very rapidly to the surface, making the cathode active again. At 1200°C, this diffusion is very slow and the cathode does not activate immediately.

The diffusion of lanthanum to the LaB_6 surface, when a vacancy exists there, was investigated quantitatively as a function of the temperature. The LaB_6 surface was first made inactive by bombardment with

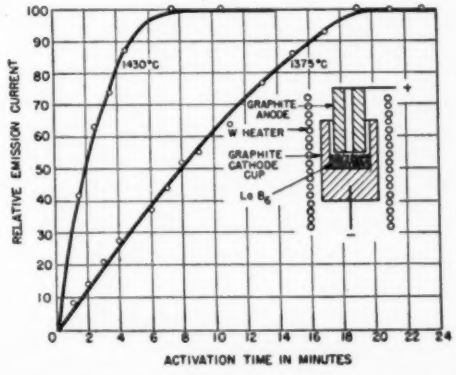


FIG. 13. Time variation of the emission at 1115°C from initially inactive lanthanum boride for two activating temperatures.

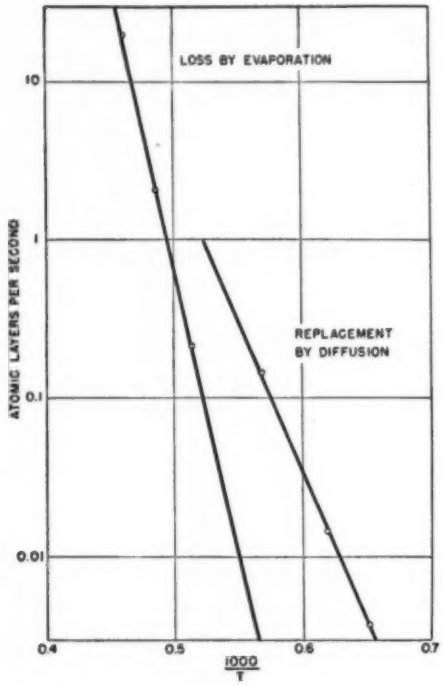


FIG. 14. Curves showing the rate at which lanthanum evaporates from the LaB_6 surface and the rate at which it is replaced by diffusion.

110-volt mercury ions. The anode voltage was then immediately reduced to approximately 18 volts and the time for the emission to recover was measured. The reciprocal of this period is a measure of the average rate of diffusion of lanthanum to the surface in monolayers per second. The results of measurements at three different temperatures are plotted in Fig. 14. Also plotted in this same figure are the evaporation rate data of Table III expressed in atomic layers per second of lanthanum rather than grams per square centimeter per second. It thus appears that in the operating temperature range, the lanthanum on the LaB_6 surface is replaced, when a vacancy occurs, at a much faster rate than the lanthanum would normally evaporate from the LaB_6. At 1500°C, for example, the average lifetime of a lanthanum atom on the LaB_6 surface is five minutes, while the average replacement time of this atom is only six seconds. At temperatures below approximately 1890°C, the evaporation rate is determined by the surface energy of the lanthanum on the LaB_6; and for temperatures above this, it is determined by the rate of diffusion of the lanthanum atoms to the surface. At current densities above 70 amperes per square centimeter, the current should increase at a rate less than that given by Eq. (1) because of lanthanum depletion on the LaB_6 surface.

The emission from the alkaline-earth borides appears to decrease with time when operated at high temperatures. It is believed that this is caused by inadequate diffusion rates of the alkaline-earth metals through the boron framework, although this was not thoroughly investigated. This seemed to be especially true for CaB_6 since a residue (consisting probably of boron) was always found on the surface after it had been operated in vacuum at high temperatures.

As would be expected from the metallic character of the boride cathodes, the pulsed emission does not differ from the steady dc emission. Figure 15 shows a Schottky plot of the emission obtained from a CeB_6 cathode. The cathode consisted of a sintered rod ground to 0.184 inch in diameter and 0.430 inch long. This was heated by the anode which consisted of a tungsten coil 0.308 inch inside diameter and 0.750 inch long. The points on this curve for field strengths below 20 kv/cm were obtained with a dc anode voltage. The points for field strengths greater than this value were obtained with 5-microsecond pulses at a 60-cycle repetition rate. It can be seen that as the electric field increases up to approximately 300 kv/cm, the slope of the curve decreases approaching the Schottky value. For fields greater than this, the slope of the curve again increases. The cause of this is uncertain. It may be field emission effects or possibly very fine scale patchiness.

Many of the LaB_6 properties already discussed have been confirmed by photoelectric measurements. A LaB_6 photo-cell was constructed as follows. A graphite cylinder was coated with LaB_6. Provisions were made for heating this cathode by radiation from a tungsten coil within. A cylindrical molybdenum anode surrounded the cathode and contained a small opening in the side for admitting the light. Corning 9741 ultraviolet glass was used in the cell envelope. The LaB_6 was sintered on the graphite at 1800°C and the anode was outgassed by electron bombardment from the cathode. Tungsten and barium getters as well as an

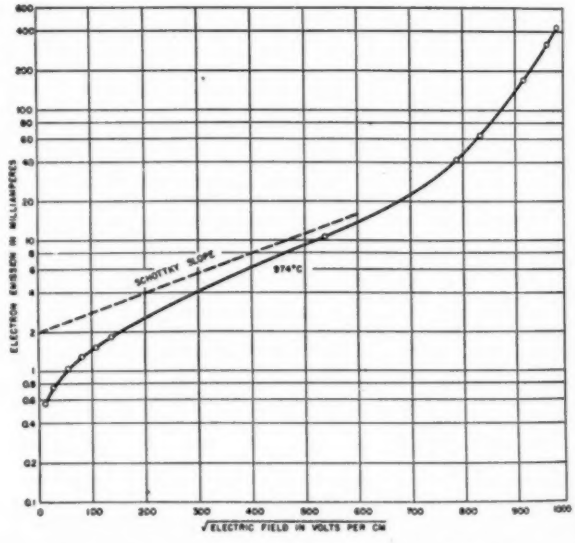


FIG. 15. A Schottky plot of the electron emission from cerium boride.

ionization gauge were sealed in appendages on the photo-cell. The vacuum was better than 10^{-7} mm of Hg after sealing off. The cathode was then flashed at 1200°C before taking data.

The spectral distribution of the photoelectric yield in electrons per quantum from LaB_6 at room temperature is shown in Fig. 16.[10] This curve is parabolic, characteristic of a metal, over at least a 0.75-electron volt range. The work function determined from this curve by Fowler's method is 2.67 volts. This in excellent agreement with 2.66 volts determined thermionically and indicates that the work function is not appreciably temperature dependent. This conclusion was checked further by varying the photo-cathode temperature from 300°K to 700°K and observing that the photoelectric yield and work function did not change.

Explorations made over the surface of the LaB_6 showed some large scale patchiness with the work function increasing by not more than 0.1 volt. Increasing the anode voltage from 10 volts to 200 volts caused the emission to increase 30 percent at the 3.39-electron volt point, indicating small scale patchiness. This is also evident for CeB_6 in Fig. 15 by the large slope of the Schottky curve at low fields.

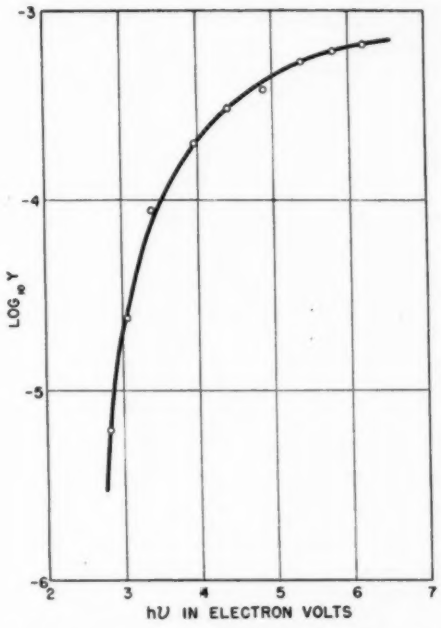


FIG. 16. Spectral distribution of the photoelectric yield in electrons per quantum from LaB_6 at 300°K. The work function determined by Fowler's method from this curve is 2.67 volts.

VIII. ACKNOWLEDGMENTS

This research was supported in part by the ONR, Contract No. NSOri-178-Task I. The writer is indebted to Dr. A. W. Hull and Dr. Saul Dushman for many stimulating discussions on various matters associated with this paper. He is also indebted to Mr. F. C. Kelley and Mr. W. J. Dondalski for their assistance in preparing and sintering the boride samples. Dr. D. Harker and his group identified many of the pure borides and solid solutions of mixed borides by x-ray analysis. Dr. H. Liebhafsky and Mr. L. B. Bronk made chemical analysis on some of the samples, and Mrs. A. P. O'Hara made spectrographic analyses of the materials used and of many of the boride products. The writer is also very grateful to Mr. C. Bell for his valuable help in preparing cathodes and expediting tube construction.

[10] The writer is indebted to Dr. L. Apker and Mr. E. Taft of this laboratory for the photoelectric data presented in this paper.

Transmission-Line Equivalent of Electronic Traveling-Wave Systems

W. E. MATHEWS
*Bell Telephone Laboratories, Murray Hill, New Jersey**
(Received August 28, 1950)

It is well known that the small-signal behavior of long electron beams may be analyzed in terms of propagating space-charge waves, suggesting an equivalence between such beams and longitudinally moving transmission lines. This in turn suggests the analysis of such electronic devices as the traveling-wave amplifier, double-stream or electron-wave amplifier, and multicavity magnetron, in terms of coupled distributed-parameter transmission lines moving relative to each other. It is shown that this approach is equivalent to a rigorous field-theory analysis in certain cases of particular interest, and the procedure for calculating the significant distributed parameters is indicated. Final results for the idealized helix and thin cylindrical electron beam are presented.

I. INTRODUCTION

HAHN[1,2] and Ramo[3,4] have shown that the small-signal behavior of long electron beams, as used in velocity modulation tubes and other modern microwave devices, may conveniently be analyzed in terms of propagating space-charge waves. They find that the ordinary high voltage cylindrical beam is capable of supporting two fundamental waves and an infinite series of higher mode pairs, one wave of each pair traveling slightly faster and the other the same amount slower than the average beam velocity. Tonks and Langmuir also have shown that such space-charge waves may exist, propagating either forward or backward with the same characteristic velocity, in a stationary electron plasma, when the plasma is excited at a frequency greater than the natural plasma resonance frequency.[5] Thus, a stationary plasma behaves, for small signals, as a high pass transmission line; and a high voltage beam is equivalent to such a line moving at the average beam velocity.

This equivalence suggests the analysis of such electronic devices as the traveling-wave amplifier, double-stream or electron-wave amplifier, and multicavity magnetron, in terms of coupled distributed-parameter transmission lines moving relative to each other. The basic theory of interaction between such lines has been worked out, and has been shown to yield results consistent with published electron-ballistic traveling-wave theory.[6] It is the purpose of the present paper to show that this approach is equivalent to a rigorous field-theory analysis in certain cases of particular interest, and to indicate the procedure for calculating the significant distributed parameters.

II. FUNDAMENTAL ASSUMPTIONS

We shall consider parallel plane and coaxial cylindrical systems, taking x as the normal coordinate in the plane case. We shall assume that motion of the electrons is confined to the z direction and that all signal components vary as $e^{j(\omega t-\beta z)}$. We also shall assume that the electron velocity is a small fraction of the velocity of light, so that nonrelativistic equations apply and so that the longitudinal force due to magnetic fields may be neglected in comparison with the force due to the longitudinal electric field E_z. We shall restrict the analysis to small signals, so that cross products of alternating current quantities may be neglected. Finally, we shall consider only laterally invariant ($\partial/\partial y=0$ in plane case) or circularly symmetric ($\partial/\partial\varphi=0$ in cylindrical case) modes.

III. THE PROBLEM IN TERMS OF FIELD THEORY

From the point of view of field theory, analysis of the behavior of any electromagnetic system or device involves (a) dividing the system into convenient regions, for each of which the field equations may be formulated, (b) setting up the general solution of the field equations for each region, and (c) matching these solutions simultaneously at all the boundaries between regions. The helix traveling-wave tube, for example, naturally divides into (1) the region outside the helix, (2) the region between the helix and the electron beam, and (3) the region occupied by the beam. The field equations for each of these regions may be solved readily, and the crux of the problem is to match these solutions at the helix and beam surfaces. Such analyses of the traveling-wave tube have been made by Chu and Jackson[7] and Rydbeck.[8]

Fundamentally, the field-matching requirement at a mathematical (zero thickness) boundary is that the tangential components of E and H must be continuous across the boundary. It is convenient to generalize the boundary concept to include thin elements capable of supporting finite currents, such as a conducting sheet or thin hollow-cylindrical electron beam, assuming as the criterion of thinness that E is essentially constant through the element. The generalized requirement on

* Now at California Institute of Technology, Pasadena, California.

[1] W. C. Hahn, Gen. Elec. Rev. **42**, 258–270 (June, 1939).
[2] W. C. Hahn, Gen. Elec. Rev. **42**, 497–502 (November, 1939).
[3] S. Ramo, Phys. Rev. **56**, 276–283 (1939).
[4] S. Ramo, Proc. Inst. Radio Engrs. **27**, 757–763 (1939).
[5] L. Tonks and I. Langmuir, Phys. Rev. **33**, 195–210 (1929).
[6] W. E. Mathews, Proc. Inst. Radio Engrs. (to be published).
[7] L. J. Chu and D. Jackson, "Field theory of traveling-wave tubes," Technical Report No. 38, Research Laboratory of Electronics, M.I.T., April, 1947. Also, Proc. Inst. Radio Engrs. **36**, 853–863 (1948).
[8] O. E. H. Rydbeck, "The theory of the traveling-wave tube," Ericsson Technics No. 46 (1948).

H, therefore, is that the difference between the tangential components of H at the two surfaces of the boundary element must be equal to the density of current (current per unit width) flowing in the element.[9] Thus, if the element lies in the yz plane, we may write

$$H_{y2}-H_{y1}=J_z, \tag{1}$$

$$-H_{z2}+H_{z1}=J_y, \tag{2}$$

where the subscripts 1 and 2 refer to the surfaces of lesser and greater x, respectively, and J_z and J_y are the components of total current per unit y and z, respectively. If the element is cylindrical, the H matching requirement becomes

$$H_{\varphi 2}-H_{\varphi 1}=J_z, \tag{3}$$

$$-H_{z2}+H_{z1}=J_\varphi, \tag{4}$$

where the subscripts 1 and 2 now refer to the surfaces of lesser and greater r, respectively, and J_z and J_φ are the components of total current per unit $r\varphi$ and z, respectively.

Since E is assumed constant through the generalized boundary, we may divide Eqs. (1) through (4) by appropriate components of E to obtain

$$\begin{aligned} H_{y2}/E_z-H_{y1}/E_z&=J_z/E_z,\\ H_{\varphi 2}/E_z-H_{\varphi 1}/E_z&=J_z/E_z, \end{aligned} \tag{5}$$

$$\begin{aligned} -H_{z2}/E_y+H_{z1}/E_y&=J_y/E_y,\\ -H_{z2}/E_\varphi+H_{z1}/E_\varphi&=J_\varphi/E_\varphi. \end{aligned} \tag{6}$$

Taking z as the longitudinal direction, and x as the coordinate perpendicular to the planes in the parallel-plane case, we introduce the assumption $\partial/\partial y=0$ (no transverse field variation parallel to the planes) or $\partial/\partial\varphi=0$ (circularly symmetric fields) as appropriate. We then find that the general solution of the field equations for a homogeneous isotropic region reduces to two independent components—a TM component involving E_z, E_x, and H_y (or E_z, E_r, and H_φ) and a TE component involving H_z, H_x, and E_y (or H_z, H_r, and E_φ). Thus, H_y/E_x (or H_φ/E_r) and $-H_z/E_y$ (or $-H_z/E_\varphi$) are seen to be the transverse admittance functions, measured in the direction of decreasing x (or r), for TM and TE waves, respectively.[10] Defining Y_{t1} and Y_{t2} as the transverse admittances looking *away* from the boundary element at surfaces 1 and 2, respectively, and Y_e as the corresponding surface admittance (J/E) of the element itself, we may express both Eq. (5) and Eq. (6) in the form

$$Y_{t1}+Y_{t2}+Y_e=0. \tag{7}$$

We note that in general, of course, Eq. (7) must be satisfied independently for the TM and TE components.

Restricting our attention to systems in which at least one of the interacting elements is an electron beam, we assume that motion of the electrons is confined to the longitudinal direction, and that the longitudinal force due to magnetic fields can be neglected in comparison with the force due to the electric field. This implies that the electron beam is affected only by the longitudinal component of electric field E_z, and thus that the beam and circuit are coupled solely through the TM component of the total wave. As Chu and Jackson have pointed out, the boundary-value problem in this case reduces to a matching of TM transverse admittances.[7] If the circuit is capable of propagating a purely TM wave, there need be no TE component at all in the coupled system. If, on the other hand, the circuit inherently requires a TE component, as does the helix, this component still does not enter explicitly in the boundary-matching relations, but rather appears implicitly in the TM surface admittance of the circuit.

IV. EQUIVALENT LUMPED-ELEMENT CIRCUIT

The field-matching requirement at any boundary, as given by Eq. (7), is that the sum of the admittances vanish. L. R. Walker has pointed out, however, that this is precisely what Kirchoff's laws require of a parallel-connected network in which currents are flowing. Now, the space between any two boundaries may be viewed, from the standpoint of transverse admittances, as a transmission-line transformer characteristic of the geometry of the space and the material filling it. Like any other transducer, therefore, this space may be represented by an equivalent T or π network. Thus, the transverse admittance relations for any longitudinally uniform transmission system involving field variations with only one transverse coordinate, and requiring explicit matching of only the TM or only the TE component, are formally identical with the admittance relations for an appropriate single lumped-element circuit.

The basic form of this circuit for a two-element system is shown in Fig. 1, where Y_1 represents one interacting element (e.g., a helix), Y_2 represents the other element (e.g., an electron beam), and the network Y_a, Y_b, Y_c represents the intervening space. Y_1 and Y_2 are to be measured looking toward the respective elements they represent, and in each case at the surface nearest the other element. Thus, in the case of the ordinary traveling-wave tube, Y_1 would be the admittance looking outward at the inside surface of the helix, while Y_2 would be the admittance looking into the beam at its surface. In the notation of Eq. (7), Y_1 would correspond to Y_e+Y_{t2} at the helix, while Y_{in} would correspond to Y_{t1} at the helix. Likewise, Y_2 would correspond to Y_c+Y_{t1} at the boundary of the beam, Y_e being equal to J_z/E_z for a thin hollow beam, and equal to zero for a thick solid beam.

The circuit of Fig. 2 is obviously equivalent to that of Fig. 1. With Y_1 and Y_2, however, we arbitrarily have associated Y_{s1} and Y_{s2}, which are the free-space admittances which would be seen looking toward the

[9] S. A. Schelkunoff, *Electromagnetic Waves* (D. Van Nostrand Company, Inc., New York, 1943), p. 74.

[10] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941), p. 354. The term "transverse" refers to the direction of the associated Poynting vector.

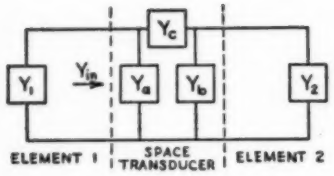


FIG. 1. Basic circuit equivalent to two-element longitudinally uniform transmission system, coupled through only *TM* or *TE* (not both) component of field.

position of the other element if that element were absent. Thus, the combination Y_1 and Y_{s1} represents element 1 in free space, with which we associate the conventional transmission-line parameters L_1 and C_1 (or v_1 and Z_{01}); and, likewise, Y_2+Y_{s2} represents element 2 in free space. The derived transducer Y_a', Y_b', Y_c, like the original transducer Y_a, Y_b, Y_c, is independent of Y_1 and Y_2, and thus depends only on the geometry of the intervening space.[11]

If one or both of the interacting elements are thin compared to a wavelength, the circuit of Fig. 2 may usefully be expanded still further. Thus, Fig. 3 might represent a traveling-wave tube with a thin hollow cylindrical beam. In this circuit we have divided Y_1 and Y_2 into two components each, one representing the admittance of free space looking away from the other element, and the other representing the effectively paralleled surface admittance of the element itself.

The admittances which we have been discussing above are basically field admittances, having the dimensions of amperes per unit width/volts per unit length. Thus, we may think of the equivalent circuit as representing a section of unit length (assumed small compared to a wavelength) of the traveling-wave system, and of the various Y's as the admittances per unit width (measured circumferentially in the cylindrical case) of this section. For formal identification with circuit equations we may consider the "voltage" across Y_1' (see Fig. 3) to be $E_{z1}=j\beta e_1$, where e_1 is the potential at the position of element 1, and likewise the "voltage" across Y_2' to be $E_{z2}=j\beta e_2$. Then the "current" through Y_1' is J_{z1}, the longitudinal current per unit width in element 1; and the "current" through Y_2' is J_{z2}.

Parallel Plane Case

Taking x as the coordinate perpendicular to the planes in the parallel plane case, we may define the transverse admittance for *TM* waves by

$$Y=H_y/E_z. \tag{8}$$

Thus defined, the admittance is to be measured in the direction of decreasing x. The transverse propagation constant γ, assumed to have no y component, is given by

$$\gamma=[\beta^2-(\omega/c)^2]^{\frac{1}{2}}\approx(\beta^2)^{\frac{1}{2}}. \tag{9}$$

Now in the charge-free nonconducting regions

$$H_y=(j\omega\epsilon/\gamma^2)(\partial E_z/\partial x), \tag{10}$$

and the field in these regions must be of the form

$$E_z=Ae^{\gamma x}+Be^{-\gamma x}. \tag{11}$$

Therefore,

$$Y=j\omega\epsilon(Ae^{\gamma x}-Be^{-\gamma x})/\gamma(Ae^{\gamma x}+Be^{-\gamma x}). \tag{12}$$

Looking into free space in the direction of decreasing x, the constant B must be identically zero. Thus, the admittance Y_s of free space is given by

$$Y_s=j\omega\epsilon/\gamma. \tag{13}$$

In the region between the interacting elements, however, both A and B may have finite values. For this region, we deduce from Eq. (12) that

$$(Y-Y_s)e^{2\gamma x}/(Y+Y_s)=\text{constant}. \tag{14}$$

In particular, if we assume x to increase in going from element 2 to element 1, and designate by Y_{in} the admittance looking toward element 2 from the inside surface of element 1, we may write

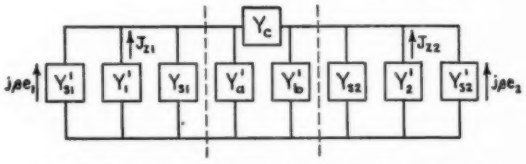


FIG. 3. Equivalent circuit for transmission system involving thin interacting elements.

$$(Y_{in}-Y_s)/(Y_{in}+Y_s)=(Y_2-Y_s)e^{-2\Delta}/(Y_2+Y_s), \tag{15}$$

$$\Delta=\gamma(x_1-x_2)>0. \tag{16}$$

Referring to Fig. 1 and making use of Eq. (15), if $Y_2=\infty$,

$$Y_{in}(\infty)=Y_a+Y_c=Y_s(1+e^{-2\Delta})/(1-e^{-2\Delta})=Y_s\coth\Delta, \tag{17}$$

while if $Y_2=0$,

$$Y_{in}(0)=Y_a+Y_bY_c/(Y_b+Y_c)=Y_s(1-e^{-2\Delta})/(1+e^{-2\Delta})=Y_s\tanh\Delta. \tag{18}$$

Recognizing that symmetry requires Y_a and Y_b to be equal, we may solve Eqs. (17) and (18) simultaneously to yield

$$Y_a=Y_b=Y_s\tanh(\Delta/2), \tag{19}$$

$$Y_c=Y_s\,\text{csch}\Delta. \tag{20}$$

Coaxial Cylindrical Case

We define the transverse admittance for circularly symmetric *TM* waves by

$$Y=H_\varphi/E_z. \tag{21}$$

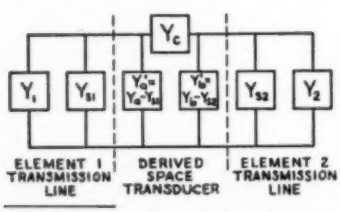


FIG. 2. Equivalent circuit in which appropriate free-space admittances are associated with the interacting elements to represent the equivalent (uncoupled) transmission lines.

[11] The transducer parameters generally are, however, functions of the absolute value of the propagation constant.

Thus defined, the admittance is to be measured in the direction of decreasing r. Equation (9) remains valid, and Eqs. (10) and (11) become[7]

$$H_\varphi=(j\omega\epsilon/\gamma)[CI_1(\gamma r)-DK_1(\gamma r)], \tag{22}$$

$$E_z=CI_0(\gamma r)+DK_0(\gamma r), \tag{23}$$

so that

$$Y=Y_s\frac{CI_1(\gamma r)-DK_1(\gamma r)}{CI_0(\gamma r)+DK_0(\gamma r)}, \tag{24}$$

where Y_s is defined by Eq. (13).

Looking into free space in the direction of decreasing r, the constant D must be identically zero. Thus, the admittance seen in this case is

$$(Y_{sr})_{\text{in}}=Y_s[I_1(\gamma r)/I_0(\gamma r)]. \tag{25}$$

Looking into free space in the direction of increasing r, the constant C must be identically zero and the admittance seen will be the negative of Y as defined by Eq. (21). Thus,

$$(Y_{sr})_{\text{out}}=Y_s[K_1(\gamma r)/K_0(\gamma r)]. \tag{26}$$

In the region between the coaxial interacting elements, both C and D may have finite values. For this region, we deduce from Eq. (24) that

$$\frac{YI_0(\gamma r)-Y_sI_1(\gamma r)}{YK_0(\gamma r)+Y_sK_1(\gamma r)}=\text{constant}. \tag{27}$$

In particular, if we assume r to increase in going from element 2 to element 1, and designate Y_{in} as before, we may write

$$\frac{Y_{\text{in}}I_0(\gamma a)-Y_sI_1(\gamma a)}{Y_{\text{in}}K_0(\gamma a)+Y_sK_1(\gamma a)}=\frac{Y_2I_0(\gamma b)-Y_sI_1(\gamma b)}{Y_2K_0(\gamma b)+Y_sK_1(\gamma b)}, \tag{28}$$

where a is the inside radius of element 1 and b is the outside radius of element 2. Successively setting $Y_2=\infty$, $Y_2=0$, and $Y_{\text{in}}=\infty$, we obtain from Eq. (28) three equations which may be solved simultaneously to yield, with reference to Fig. 1,

$$Y_a=Y_s\frac{I_1(\gamma a)K_0(\gamma b)+K_1(\gamma a)I_0(\gamma b)}{I_0(\gamma a)K_0(\gamma b)-K_0(\gamma a)I_0(\gamma b)}-Y_c, \tag{29}$$

$$Y_b=Y_s\frac{I_0(\gamma a)K_1(\gamma b)+K_0(\gamma a)I_1(\gamma b)}{I_0(\gamma a)K_0(\gamma b)-K_0(\gamma a)I_0(\gamma b)}-Y_c, \tag{30}$$

$$Y_c=Y_s/\gamma(ab)^{\frac{1}{2}}[I_0(\gamma a)K_0(\gamma b)-K_0(\gamma a)I_0(\gamma b)]. \tag{31}$$

Approximate Equivalent Circuit for Small Coupling

Small coupling between the interacting elements implies a relatively large Δ in the parallel plane case, or a γa considerably larger than γb in the coaxial cylindrical case. Noting that, for large Δ,

$$\tanh(\Delta/2)=(1-e^{-\Delta})/(1+e^{-\Delta})\approx 1-2e^{-\Delta}\approx 1-\text{csch}\Delta, \tag{32}$$

we may rewrite Eqs. (19) and (20) as

$$Y_a=Y_b\approx Y_s-Y_c, \tag{33}$$

$$Y_c=Y_s\,\text{csch}\Delta\approx Y_s2e^{-\Delta}. \tag{34}$$

In the cylindrical case we note that, for large γa,

$$I_0(\gamma a)/K_0(\gamma a)\gg 1,\quad I_1(\gamma a)/K_0(\gamma a)\gg 1,\quad K_1(\gamma a)/K_0(\gamma a)=0(1). \tag{35}$$

By virtue of Eq. (35), we may approximate Eqs. (29) through (31) by

$$Y_a\approx Y_s[I_1(\gamma a)/I_0(\gamma a)]-Y_c=(Y_{sa})_{\text{in}}-Y_c, \tag{36}$$

$$Y_b\approx Y_s[K_1(\gamma b)/K_0(\gamma b)]-Y_c=(Y_{sb})_{\text{out}}-Y_c, \tag{37}$$

$$Y_c\approx Y_s/\gamma(ab)^{\frac{1}{2}}I_0(\gamma a)K_0(\gamma b). \tag{38}$$

We see immediately from Eqs. (33), (36), and (37) that, for small coupling, Y_a' and Y_b' of the equivalent circuits of Figs. 2 and 3 are given approximately by

$$Y_a'\approx Y_b'\approx -Y_c. \tag{39}$$

Thus we may assign values to the elements of Fig. 3 as indicated in Fig. 4.

V. DISTRIBUTED-PARAMETER LINE FORMULATION

Taking "voltages" and "currents" as shown in Fig. 3 and discussed in the preceding section, we may describe the behavior of the equivalent circuit by the equations

$$j\beta(Y_{s1}'+Y_{s1}+Y_a'+Y_c)e_1-j\beta Y_ce_2=-J_{s1}, \tag{40}$$

$$j\beta Y_1'e_1=J_{s1}, \tag{41}$$

$$j\beta(Y_{s2}'+Y_{s2}+Y_b'+Y_c)e_2-j\beta Y_ce_1=-J_{s2}, \tag{42}$$

$$j\beta Y_2'e_2=J_{s2}. \tag{43}$$

Now for the slow waves which we are considering, Y_a', Y_b', Y_c, and the four free-space admittances are all essentially purely capacitive. If the two elements of the system are independently capable of wave propagation, therefore, Y_1' and Y_2' must be essentially inductive. Assuming that Y_1' and Y_2' are *purely* inductive, which is nearly true for the helix, as used in the traveling-wave tube, and for small or thin electron beams, we may re-

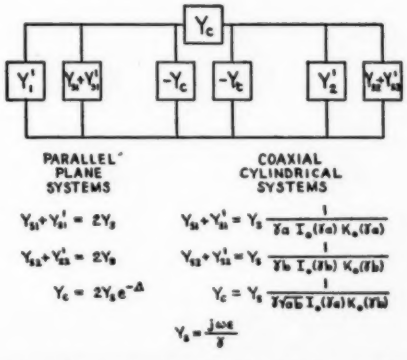


FIG. 4. Approximate equivalent circuit for small coupling.

write Eqs. (40) through (43) in the form

$$C_1'(\partial e_1/\partial t)+C_m'(\partial e_2/\partial t)=-\partial J z_1/\partial z, \quad (44)$$

$$L_1'(\partial J_{s1}/\partial t)=-\partial e_1/\partial z, \quad (45)$$

$$C_2'(\partial e_2/\partial t)+C_m'(\partial e_1/\partial t)=-\partial J_{s2}/\partial z, \quad (46)$$

$$L_2'(\partial J_{s2}/\partial t)=-\partial e_2/\partial z, \quad (47)$$

where

$$C_1'=(\beta^2/j\omega)(Y_{s1}'+Y_{s1}+Y_a'+Y_c), \quad (48)$$

$$C_2'=(\beta^2/j\omega)(Y_{s2}'+Y_{s2}+Y_b'+Y_c), \quad (49)$$

$$C_m'=-(\beta^2/j\omega)Y_c, \quad (50)$$

$$L_1'=1/j\omega Y_1', \quad (51)$$

$$L_2'=1/j\omega Y_2'. \quad (52)$$

Equations (44) through (47) clearly are formally identical with the equations for two capacitively-coupled parallel-plane transmission lines.

In the coaxial cylindrical case, the current densities J_{s1} and J_{s2} may be replaced by the total conduction currents i_1 and i_2^* by means of the relations

$$i_1=2\pi a J_{s1}, \quad (53)$$

$$i_2^*=2\pi b J_{s2}. \quad (54)$$

Equations (44) through (47) then become

$$C_1(\partial e_1/\partial t)+(a/b)^{\frac{1}{2}}C_m(\partial e_2/\partial t)=-\partial i_1/\partial z, \quad (55)$$

$$L_1(\partial i_1/\partial t)=-\partial e_1/\partial z, \quad (56)$$

$$C_2(\partial e_2/\partial t)+(b/a)^{\frac{1}{2}}C_m(\partial e_1/\partial t)=-\partial i_2^*/\partial z, \quad (57)$$

$$L_2^*(\partial i_2^*/\partial t)=-\partial e_2/\partial z, \quad (58)$$

where

$$C_1=2\pi a C_1', \quad (59)$$

$$C_2=2\pi b C_2', \quad (60)$$

$$C_m=2\pi(ab)^{\frac{1}{2}}C_m', \quad (61)$$

$$L_1=L_1'/2\pi a, \quad (62)$$

$$L_2^*=L_2'/2\pi b. \quad (63)$$

Whereas we most commonly formulate the equations for any transmission line in terms of the current, potential, and parameters that would be measured by an observer stationary with respect to that line, it must be recognized that Eqs. (44) through (47) and (55) through (58) describe the behavior of the two lines as seen from a single frame of reference. Now, if there is relative motion between the lines, as there is in all active traveling-wave devices, this frame may be stationary with respect to at most one of the lines. Let us choose a reference frame stationary with respect to line 1, and assume line 2 to be moving in the positive z direction with a uniform velocity v with respect to this frame.

Since current is defined by a single differentiation of the instantaneous charge distribution with respect to time, the line current i_2, measured in a reference frame moving with line 2, is related to i_2^* by

$$i_2=\omega_2 i_2^*/\omega, \quad (64)$$

where ω_2, the apparent frequency in the moving frame, is given by

$$\omega_2=\omega-\beta v. \quad (65)$$

The instantaneous potential e_2, on the other hand, is very nearly independent of motion of the frame in which it is measured, for velocities slow compared with the velocity of light. Thus the coefficients of Eqs. (46) and (57) are nearly independent of reference frame motion, while the coefficients of Eqs. (47) and (58) must be of such form as to make the $L(\partial i/\partial t)$ terms invariant with change of frame. Thus L_2, the inductance per unit length of line 2 measured in the moving frame, is given by

$$j\omega_2 L_2 i_2=j\omega L_2^* i_2^*, \quad (66)$$

$$L_2=\omega^2 L_2^*/\omega_2^2. \quad (67)$$

Using the subscript 2 to designate differentiation in the moving frame, we now may rewrite Eqs. (57) and (58) in terms of i_2 and L_2:

$$C_2(\partial e_2/\partial t_2)+(b/a)^{\frac{1}{2}}C_m(\partial e_1/\partial t_2)=-\partial i_2/\partial z_2, \quad (68)$$

$$L_2(\partial i_2/\partial t_2)=-\partial e_2/\partial z_2. \quad (69)$$

Small-Coupling Coaxial Cylindrical Case

The final equations for the coaxial cylindrical case are Eqs. (55), (56), (68), and (69). Combining the various steps and assuming small coupling so that Eq. (39) applies, we may summarize the parametric relationships as follows:

$$C_1=2\pi\epsilon\gamma a\{[I_1(\gamma a)/I_0(\gamma a)]+[K_1(\gamma a)/K_0(\gamma a)]\}=2\pi\epsilon/I_0(\gamma a)K_0(\gamma a), \quad (70)$$

$$C_2=2\pi\epsilon\gamma b\{[I_1(\gamma b)/I_0(\gamma b)]+[K_1(\gamma b)/K_0(\gamma b)]\}=2\pi\epsilon/I_0(\gamma b)K_0(\gamma b), \quad (71)$$

$$C_m=-2\pi\epsilon/I_0(\gamma a)K_0(\gamma b), \quad (72)$$

$$L_1=1/2\pi a j\omega Y_1', \quad (73)$$

$$L_2=\omega/2\pi b j\omega_2^2 Y_2'. \quad (74)$$

The factor $[I_0(x)K_0(x)]^{-1}$ of Eqs. (70) and (71) is plotted against x in Fig. 5. It will be seen that this function is approximated closely by the asymptotic value $2x$ for values of x greater than about 0.4.

It will be noted that C_1 and L_1 characterize element 1 in free space, and that C_2 and L_2 characterize element 2 in free space. Thus, instead of calculating these parameters theoretically, we might determine them from experimental observations of the individual elements. In particular, if we can determine the natural phase velocity v_p and characteristic impedance Z_0 of either element, the parameters L and C for that element are

given by the well-known relations

$$L=Z_0/v_p, \tag{75}$$

$$C=1/Z_0v_p. \tag{76}$$

VI. SURFACE ADMITTANCE CALCULATIONS

The preceding discussion has purposely been kept as general as possible, without restriction as to the exact nature of either of the interacting elements. By way of example, however, and as an aid in application to the traveling-wave tube, the surface-admittance calculations for the idealized helical sheet and the thin cylindrical electron beam are outlined in this section.

In accordance with the definitions (8) and (21), the transverse admittance to *TM* waves Y' of a thin (compared to a wavelength) longitudinal conducting element is the change in Y due to conduction current observed in traversing the element in the direction of increasing x or r. Since E_z must be continuous and approximately constant through a thin element, Y' is given by

$$Y'=\Delta H_y/E_z \quad \text{or} \quad Y'=\Delta H_\varphi/E_z, \tag{77}$$

which imply by Ampere's law, for either rectangular or cylindrical coordinates, that

$$Y'=J_z/E_z, \tag{78}$$

where J_z is the longitudinal conduction current per unit width (y or φ direction). Obviously, Eq. (78) is equally valid for an element such as a thin electron beam, in which case J_z represents convection current.

Idealized Helical Sheet

The field components associated with an idealized cylindrical surface of radius a, conducting in a helical direction making an angle ψ with the circumference and nonconducting in the helical direction normal to this, have been calculated by Pierce[12] and by Chu and Jackson.[7] Using Pierce's notation, the tangential components of magnetic field normal to the helical direction (H_{n1} just inside the helix and H_{n2} just outside) are given by

$$H_{n1}=H_{z1}\cos\psi-H_{\varphi1}\sin\psi=B_1I_0(\gamma a)\cos\psi - B_3(j\omega\epsilon/\gamma)I_1(\gamma a)\sin\psi, \tag{79}$$

$$H_{n2}=H_{z2}\cos\psi-H_{\varphi2}\sin\psi=B_2K_0(\gamma a)\cos\psi + B_4(j\omega\epsilon/\gamma)K_1(\gamma a)\sin\psi, \tag{80}$$

where

$$B_1=\gamma I_0(\gamma a)\tan\psi B_3/j\omega\mu I_1(\gamma a), \tag{81}$$

$$B_2=-\gamma I_0(\gamma a)\tan\psi B_3/j\omega\mu K_1(\gamma a), \tag{82}$$

$$B_4=I_0(\gamma a)B_3/K_0(\gamma a). \tag{83}$$

Now

$$J_z=J\sin\psi=(H_{n1}-H_{n2})\sin\psi, \tag{84}$$

and

$$E_z=B_3I_0(\gamma a). \tag{85}$$

[12] J. R. Pierce, Proc. Inst. Radio Engrs. **35**, 111–123 (1947).

Combining Eqs. (79) through (85) we obtain

$$Y_{\text{helix}}'=\frac{J_z}{E_z}=\sin^2\psi\left\{\frac{\gamma}{j\omega\mu}\left[\frac{I_0(\gamma a)}{I_1(\gamma a)}+\frac{K_0(\gamma a)}{K_1(\gamma a)}\right] - \frac{j\omega\epsilon}{\gamma}\left[\frac{I_1(\gamma a)}{I_0(\gamma a)}+\frac{K_1(\gamma a)}{K_0(\gamma a)}\right]\right\}$$

$$=\frac{\sin^2\psi}{\gamma a}\left[\frac{\gamma}{j\omega\mu I_1(\gamma a)K_1(\gamma a)}-\frac{j\omega\epsilon}{\gamma I_0(\gamma a)K_0(\gamma a)}\right]. \tag{86}$$

We see from Eq. (86) that the helical surface behaves as a distributed inductance in parallel with a distributed negative capacitance. For the slow waves in which we are interested, however,

$$\omega\epsilon/\gamma\ll\gamma/\omega\mu, \tag{87}$$

so that Eq. (86) becomes approximately

$$Y_{\text{helix}}'\approx\sin^2\psi/j\omega\mu aI_1(\gamma a)K_1(\gamma a). \tag{88}$$

To determine the propagation behavior of the helix in free space, we write the admittance-matching equation

$$Y_{\text{helix}}'+(Y_{sa})_{\text{in}}+(Y_{sa})_{\text{out}}=0, \tag{89}$$

where the three Y's are given by Eqs. (86), (25), and (26), respectively. This yields immediately

$$I_1(\gamma a)K_1(\gamma a)\beta_0^2\cot^2\psi/I_0(\gamma a)K_0(\gamma a)\gamma^2=1, \tag{90}$$

which is identical with Pierce's Eq. (130) and Chu and Jackson's Eq. (38). Use of Eq. (88) instead of Eq. (86) in Eq. (89) yields

$$I_1(\gamma a)K_1(\gamma a)\beta_0^2\csc^2\psi/I_0(\gamma a)K_0(\gamma a)\gamma^2=1. \tag{91}$$

The factors $[I_0(x)K_0(x)]^{-1}$ and $[I_1(x)K_1(x)]^{-1}$ of Eqs. (86) through (91) are plotted against x in Fig. 5. It will be seen that both of these functions are approximated closely by the asymptotic value $2x$ for values of x greater than about 3. Thus, for $\gamma a>3$ the helix is essentially nondispersive (γ independent of γa), and Eq. (88)

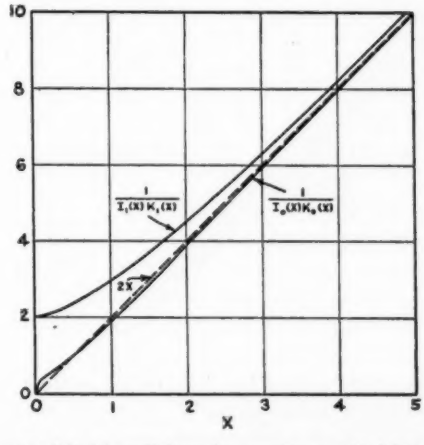


FIG. 5. Modified bessel function products $1/I_0(x)K_0(x)$ and $1/I_1(x)K_1(x)$ *versus* x.

becomes

$$Y_{\text{helix}}' \approx 2\gamma \sin^2\psi/j\omega\mu. \quad (92)$$

Substituting Eqs. (88) and (92) into Eq. (73), we find the equivalent transmission-line inductance for the helix to be

$$L_1 = \mu I_1(\gamma a)K_1(\gamma a)/2\pi \sin^2\psi \approx \mu/4\pi\gamma a \sin^2\psi, \quad (93)$$

the latter expression being valid when $\gamma a > 3$.

Thin Electron Beam

The small-signal surface admittance of a thin electron beam may be calculated immediately from the electron equation of motion

$$v_e = j\eta E_z/(\omega - \beta v_0), \quad (94)$$

the continuity equation

$$j\beta J_z = j\omega\sigma, \quad (95)$$

and the first-order convection current definition

$$J_z = \sigma v_0 + \sigma_0 v_e, \quad (96)$$

where η is the charge-to-mass ratio of the electron, v_0 the average beam velocity, σ_0 the average surface charge density, and v_e and σ the alternating components of velocity and charge density, respectively. Combining these relations, we obtain

$$Y_{\text{beam}}' = J_z/E_z = j\eta\sigma_0\omega/(\omega - \beta v_0)^2. \quad (97)$$

Now from Eq. (65), the apparent signal frequency ω_2 in a reference frame moving with the beam ($v = v_0$) is given by

$$\omega_2 = \omega - \beta v_0, \quad (98)$$

so that Eq. (97) may be written

$$Y_{\text{beam}}' = j\eta\sigma_0\omega/\omega_2^2. \quad (99)$$

Substituting Eq. (99) into Eq. (74), we find the inductance per unit length of the equivalent transmission line, moving at the average beam velocity, to be

$$L_2 = -1/2\pi b\eta\sigma_0 = (2V_0/\eta)^{\frac{1}{2}}/I_0, \quad (100)$$

where I_0 and V_0 are the total beam current and accelerating voltage, respectively. Being independent of b, the beam radius, Eq. (100) applies to thin solid as well as hollow beams.

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

Thickness-Shear and Flexural Vibrations of Crystal Plates

RAYMOND D. MINDLIN
Columbia University, New York, and Bell Telephone Laboratories, Murray Hill, New Jersey

(Received September 18, 1950)

The theory of flexural motions of elastic plates, including the effects of rotatory inertia and shear, is extended to crystal plates. The equations are solved approximately for the case of rectangular plates excited by thickness-shear deformation parallel to one edge. Results of computations of resonant frequencies of rectangular, *AT*-cut, quartz plates are shown and compared with experimental data. Simple algebraic formulas are obtained relating frequency, dimensions, and crystal properties for resonances of special interest in design.

INTRODUCTION

EXACT solutions of the three-dimensional equations governing the vibrations of crystals have been obtained for only a few body shapes and modes of motion. Special interest centers around a few shapes and modes which have been found useful as frequency control and filter elements in electric circuits. Among these are the modes excited by forced thickness-shear deformation of plates.

The resonant frequencies of thickness-shear modes in an infinite crystal plate were found by Koga[1] in closed form, but correspondingly simple solutions of the thickness-shear type are not obtainable for a bounded plate. Using Koga's solution as a zero-order approximation, Ekstein[2] applied a perturbation method to obtain approximate frequencies of a plate with a pair of parallel bounding planes normal to the faces of the plate. He arrived at a formula for the resonant frequencies which he showed to have the same form as an empirical formula devised by Sykes.[3] This formula, however, fits Sykes' experimental data only at discrete points whose locations cannot be found from the formula. The curves of frequency *versus* length to thickness ratio obtained from Sykes' and Ekstein's formulas depart markedly from the experimental data in the neighborhood of flexural resonances and become progressively worse both at higher thickness-shear overtones and lower length-to-thickness ratios. Ekstein advanced a qualitative explanation for the discrepancies in the neighborhood of flexural resonances by pointing out that

[1] I. Koga, Physics **3**, 70 (1932).
[2] H. Ekstein, Phys. Rev. **68**, 11 (1945).
[3] R. A. Sykes, Chapter VI, *Quartz Crystals for Electrical Circuits*, edited by R. A. Heising (D. Van Nostrand, New York, 1946), p. 218, Eqs. 6.9.

flexural modes in an infinite plate constitute another zero-order solution, and he had shown previously[4] that "coupling" effects become important when the frequencies of two zero-order modes are nearly equal.

In Ekstein's method of solving plate vibration problems, a start is made with one or more exact solutions (zero-order approximations) of the three-dimensional equations, satisfying some of the boundary conditions (e.g., those on the faces of the plate), and a perturbation theory is applied to satisfy, approximately, the remaining boundary conditions. On the other hand, for bodies having at least one small dimension, the classical procedure in elasticity theory is to reduce the number of variables in the differential equations of the system at the start by omitting certain variables deemed to be unimportant and employing average values or restricted regions of others. In this way an approximate theory is formed (two- or one-dimensional) in which some of the boundary conditions are satisfied identically, and some or all of the remaining ones may be satisfied exactly. Familiar examples are the approximate theories of flexural motions of bars and plates leading to the bernoulli-euler and lagrange equations of motion. In their classical form, these theories are not suitable for the present problem as they contain no shear deformation and, as Rayleigh[5] pointed out, they give good results for flexural wave velocities only when the wavelength is long in comparison with the thickness.

In the case of bars, Rayleigh[6] added a correction term for rotatory inertia to the classical theory with some, but not much, improvement in wave velocities. It was not until Timoshenko[7] introduced a correction for transverse shear deformation that the one-dimensional beam theory was made suitable for short wavelengths and high frequencies.

In the case of isotropic plates it has been shown[8] that the theory corresponding to Timoshenko's beam theory contains the appropriate approximations to the first thickness-shear mode and all the flexure modes. This theory is extended to crystal plates in the present paper. The additional boundaries considered by Ekstein are then introduced and a transcendental frequency equation of simple form is obtained. Notwithstanding its simplicity, the frequency equation contains, in great detail, the elaborate spectrum of frequencies obtained experimentally. The zeros and infinities of the frequency equation correspond to points in the frequency spectrum of special interest in design. For these points explicit solutions are obtained in the form of a pair of algebraic formulas. One of the formulas is similar to that of Sykes and Ekstein but contains an additional term which accommodates the higher thickness-shear overtones and smaller length to thickness ratios. The second formula locates the points at which the curves, computed from the first one, cross the curves computed from the transcendental equation. Thus, both frequency and length to thickness ratio are obtained easily for the points of interest without solving the transcendental equation.

The frequency equation that is obtained is simply the one for flexural resonances according to the improved theory of plates and is quite independent of the notion of thickness shear modes which, of course, do not exist in pure form in a bounded plate. From the point of view of plate theory, shear deformation is always present in flexural motion so that flexural resonance may be excited by forcing shear deformation (piezoelectrically or electrostrictively) at the resonant frequencies of flexure. From this point of view, the resonances in a bounded plate commonly designated as thickness-shear, or thickness-shear overtones, are simply local regions in the spectrum of flexural resonances over which the frequency does not change, as rapidly as elsewhere, with change of plate dimensions.

DERIVATION OF PLATE EQUATIONS

The plate is referred to an x, y, z system of rectangular coordinates. The faces of the plate are the planes $y=\pm h/2$, and its cylindrical surfaces (edges) have their generators normal to the x, z plane. The elastic properties of the material are assumed to have monoclinic symmetry. This includes the crystal classes and cuts commonly employed as thickness-shear resonators. With this symmetry and the x-axis as the principal axis, the stress-displacement relations are (in the I.R.E. Standard notation)

$$\begin{aligned}
T_1&=c_{11}(\partial u_1/\partial x)+c_{12}(\partial u_2/\partial y)+c_{13}(\partial u_3/\partial z)\\
&\quad+c_{14}[(\partial u_3/\partial y)+(\partial u_2/\partial z)],\\
T_2&=c_{12}(\partial u_1/\partial x)+c_{22}(\partial u_2/\partial y)+c_{23}(\partial u_3/\partial z)\\
&\quad+c_{24}[(\partial u_3/\partial y)+(\partial u_2/\partial z)],\\
T_3&=c_{13}(\partial u_1/\partial x)+c_{23}(\partial u_2/\partial y)+c_{33}(\partial u_3/\partial z)\\
&\quad+c_{34}[(\partial u_3/\partial y)+(\partial u_2/\partial z)],\\
T_4&=c_{14}(\partial u_1/\partial x)+c_{24}(\partial u_2/\partial y)+c_{34}(\partial u_3/\partial z)\\
&\quad+c_{44}[(\partial u_3/\partial y)(\partial u_2/\partial z)],\\
T_5&=c_{55}[(\partial u_1/\partial z)+(\partial u_3/\partial x)]\\
&\quad+c_{56}[(\partial u_2/\partial x)+(\partial u_1/\partial y)],
\end{aligned}\tag{1}$$

and

$$T_6=c_{56}[(\partial u_1/\partial z)+(\partial u_3/\partial x)]+c_{66}[(\partial u_2/\partial x)+(\partial u_1/\partial y)].$$

In the three-dimensional theory, substitution of these expressions for the components of stress into the stress equations of motion

$$\begin{aligned}
(\partial T_1/\partial x)+(\partial T_6/\partial y)+(\partial T_5/\partial z)&=\rho(\partial^2 u_1/\partial t^2),\\
(\partial T_6/\partial x)+(\partial T_2/\partial y)+(\partial T_4/\partial z)&=\rho(\partial^2 u_2/\partial t^2),\\
\text{and}\\
(\partial T_5/\partial x)+(\partial T_4/\partial y)+(\partial T_3/\partial z)&=\rho(\partial^2 u_3/\partial t^2)
\end{aligned}\tag{2}$$

yields the differential equations governing the displacements.

[4] H. Ekstein, Phys. Rev. **66**, 108 (1944).
[5] Lord Rayleigh, Proc. London Math. Soc. **10**, 225 (1889).
[6] Lord Rayleigh, *Theory of Sound* (Macmillan and Company, London, 1926), second edition, Vol. 1, p. 258.
[7] S. P. Timoshenko, Phil. Mag. **41**, 744 (1921).
[8] R. D. Mindlin, Trans. Am. Soc. Mech. Engr., Paper No. 50-APM-22 (1950).

The corresponding two-dimensional relations of plate theory may be obtained as follows. The normal components of the tractions on planes parallel to the surfaces of the plate are assumed to contribute negligible amounts to the flexural and shear deformations of the plate. This is equivalent to assuming that

$$T_2=0 \tag{3}$$

throughout the plate. The plane faces of the plate are completely traction-free, so that, in addition,

$$T_6=T_4=0 \quad \text{on} \quad y=\pm h/2. \tag{4}$$

The displacements are assumed to have the form

$$u_1=y\psi_x(x,z,t), \quad u_2=\eta(x,z,t), \quad \text{and} \quad u_3=y\psi_z(x,z,t). \tag{5}$$

Equations (3) and (5) serve to eliminate all compressional and face-shear modes and all thickness-shear modes except the first and its overtones. There remain only the possibilities of approximations to plate flexural, torsional, and first thickness-shear modes and their overtones.

Three stress couples and two stress resultants are defined, in the usual manner, by

$$\begin{gathered}(M_x, M_z, M_{xz})=\int_{-h/2}^{h/2}(T_1, T_3, T_5)y\,dy \\ (Q_x, Q_z)=\int_{-h/2}^{h/2}(T_6, T_4)dy.\end{gathered} \tag{6}$$

By means of (3), the second of Eq. (1) is solved for $\partial u_2/\partial y$, and this expression is used to eliminate $\partial u_2/\partial y$ from the remaining five stress-displacement relations. Equations (5) are now substituted in the remaining five of (1) and integrations are performed so as to convert the left-hand sides of (1) to stress couples and stress resultants in accordance with (6). The formulas for the two stress resultants (Q_x, Q_z) are then altered by introducing a constant multiplying factor whose magnitude is to be determined later by matching an exact solution of the three-dimensional equations. This procedure converts (1) to

$$\begin{gathered}M_x=D_1(\partial\psi_x/\partial x)+D_2\partial\psi_z/\partial z, \\ M_z=D_2(\partial\psi_x/\partial x)+D_3\partial\psi_z/\partial z, \\ M_{xz}=D_5[(\partial\psi_z/\partial x)+\partial\psi_x/\partial z], \\ Q_z=D_4(\psi_z+\partial\eta/\partial z), \quad \text{and} \quad Q_x=D_6(\psi_x+\partial\eta/\partial x),\end{gathered} \tag{7}$$

where

$$\begin{gathered}D_1=(c_{11}-c_{12}^2/c_{22})h^3/12, \quad D_2=(c_{13}-c_{12}c_{23}/c_{22})h^3/12, \\ D_3=(c_{33}-c_{23}^2/c_{22})h^3/12, \quad D_4=\kappa^2(c_{44}-c_{24}^2/c_{22})h, \\ D_5=c_{55}h^3/12, \quad \text{and} \quad D_6=\kappa^2c_{66}h,\end{gathered} \tag{8}$$

in which κ^2 is the multiplying factor mentioned above. Of the 13 elastic constants, three (c_{14}, c_{34}, c_{56}) do not contribute, and the remaining ten enter in only six combinations. It may be seen that κ controls the transverse shear correction. For example, if κ is infinite the last two of (7) require

$$\psi_z=-\partial\eta/\partial z, \quad \psi_x=-\partial\eta/\partial x, \tag{9}$$

which are the relations of classical plate theory in which shear deformation is neglected.

The equations of motion (2) are converted, similarly, by integrating over the thickness of the plate after multiplying the first and third by y. If Eqs. (3)–(6) are used, the results are

$$\begin{gathered}(\partial M_x/\partial x)-Q_x+(\partial M_{xz}/\partial z)=(\rho h^3/12)(\partial^2\psi_x/\partial t^2), \\ (\partial Q_x/\partial x)+(\partial Q_z/\partial z)=\rho h(\partial^2\eta/\partial t^2),\end{gathered}$$

and

$$(\partial M_{xz}/\partial x)-Q_z+(\partial M_z/\partial z)=(\rho h^3/12)(\partial^2\psi_z/\partial t^2). \tag{10}$$

Finally, inserting (7) into (10), the plate equations of motion are

$$\begin{gathered}D_1(\partial^2\psi_x/\partial x^2)+D_5(\partial^2\psi_x/\partial z^2) \\ \quad+(D_2+D_5)(\partial^2\psi_z/\partial x\partial z)-D_6(\psi_x+\partial\eta/\partial x) \\ \quad=(\rho h^3/12)(\partial^2\psi_x/\partial t^2), \\ D_6[(\partial\psi_z/\partial x)+(\partial^2\eta/\partial x^2)] \\ \quad+D_4[(\partial\psi_z/\partial z)+(\partial^2\eta/\partial z^2)]=\rho h(\partial^2\eta/\partial t^2),\end{gathered}$$

and

$$\begin{gathered}D_3(\partial^2\psi_z/\partial z^2)+D_5(\partial^2\psi_z/\partial x^2) \\ \quad+(D_2+D_5)(\partial^2\psi_x/\partial x\partial z)-D_4(\psi_z+\partial\eta/\partial z) \\ \quad=(\rho h^3/12)(\partial^2\psi_z/\partial t^2).\end{gathered} \tag{11}$$

The right-hand sides of the first and third of (11) represent the effects of rotatory inertia. If these are omitted, and if the shear deformation terms (i.e., those containing κ) are eliminated, Eqs. (11), with the help of (9), reduce to the classical theory of orthotropic plates. The retention of the shear and rotatory inertia terms permits the application of the theory of plates to high frequency flexural modes and the accompanying thickness-shear motions.

Equations (7) and (11) replace Hooke's law and the equations of motion of the three-dimensional theory. The associated boundary conditions may be found by converting Neumann's uniqueness theorem[9] to apply to plates. This may be done by using (5) and performing the integrations with respect to y that appear in the volume and surface integrals in Neumann's theorem.[10] In terms of components referred to coordinates ν and s measured, respectively, normal to and along the boundary of the plate, the theorem reveals that, at each point of the edge of the plate, one member of each of the following three products must be specified:

$$M_\nu(\partial\psi_\nu/\partial t), \quad M_{\nu s}(\partial\psi_s/\partial t), \quad Q_\nu(\partial\eta/\partial t). \tag{12}$$

On a free edge, for example,

$$M_\nu=M_{\nu s}=Q_\nu=0. \tag{13}$$

Thus, three boundary conditions are required at each

[9] A. E. H. Love, *Theory of Elasticity* (Cambridge University Press, London, 1927), fourth edition, p. 176.

[10] The development of a plate-analog of Neumann's uniqueness theorem is described in detail in reference 8.

point on the edge, as opposed to two in the classical theory of plates. This is in agreement with E. Reissner's theory of equilibrium of plates[11] in which shear deformation is accounted for in a different manner.

THICKNESS-SHEAR VIBRATION OF AN INFINITE PLATE

The solution of the three-dimensional equations for thickness-shear vibration parallel to the x axis is obtained by setting

$$u_1=u(y)e^{i\omega t}, \quad u_2=u_3=0. \tag{14}$$

The equations of motion (2) then reduce to

$$c_{66}(\partial^2 u/\partial y^2)+\rho\omega^2 u=0.$$

Applying the boundary conditions (4), the usual result[1] is obtained for the circular frequency $\tilde{\omega}$ of the first mode:

$$\tilde{\omega}^2=\pi^2 c_{66}/\rho h^2. \tag{15}$$

The corresponding solution of the plate, Eqs. (11), is obtained by setting

$$\psi_x=B\exp(i\tilde{\omega}t), \quad \eta=\psi_z=0, \tag{16}$$

where B is a constant. The equations of motion (11) are satisfied provided

$$\tilde{\omega}^2=12\kappa^2 c_{66}/\rho h^2. \tag{17}$$

The value of the coefficient κ is obtained by equating (15) and (17), with the result

$$\kappa=\pi/(12)^{\frac{1}{2}}. \tag{18}$$

It may be seen that the first antisymmetric mode of thickness-shear motion is contained in plate theory when rotatory inertia and shear terms are included.

THICKNESS-SHEAR EXCITATION OF A RECTANGULAR PLATE

Suppose, now, that the thickness shear vibration

$$\psi_x=Be^{i\omega t}, \quad \eta=\psi_z=0 \tag{19}$$

is forced (piezoelectrically, for example) in a rectangular plate with free edges $x=\pm l/2$, $z=\pm$const. The boundary conditions to be satisfied are, from (13),

$$M_x=M_{xz}=Q_x=0 \quad \text{on} \quad x=\pm l/2 \tag{20}$$

$$M_z=M_{xz}=Q_z=0 \quad \text{on} \quad z=\pm\text{const.} \tag{21}$$

Substitution of (19) into (7) shows that all the boundary conditions are satisfied with the exception that

$$Q_x=D_6Be^{i\omega t} \quad \text{on} \quad x=\pm l/2. \tag{22}$$

Additional modes are required such that, combined with (19), the equations of motion (11) and the boundary conditions (20) and (21) will be satisfied. Since (11) are essentially the equations of flexural motion, it is apparent that the presence of boundaries results, in this case, in the excitation of flexural modes.

[11] E. Reissner, J. Appl. Mech. **12**, A-69 (1945); Quart. Appl. Math. **5**, 55 (1947).

SOLUTION NEGLECTING WIDTH

If the additional motions introduced to remove (22) are taken as independent of z, the boundary conditions (20) on $x=\pm l/2$ can be satisfied, but small stress couples and resultants will be introduced on $z=\pm$const. Extensive experiments by Sykes on rotated Y-cut quartz plates show that, over a considerable range, the resonant frequencies are practically independent of the width. With the assumption of z-independence, ψ_z disappears from the first two equations of motion (11). These become, after introducing a factor $e^{i\omega t}$, omitted hereafter,

$$\begin{aligned}(\partial^2\psi_x/\partial x^2)+(R\delta_0^4-S^{-1})\psi_x&=S^{-1}\partial\eta/\partial x\\ (\partial\psi_x/\partial x)+[(\partial^2/\partial x^2)+S\delta_0^4]\eta&=0,\end{aligned} \tag{23}$$

where

$$R=h^2/12, \quad S=D_1/D_6,$$

and

$$\delta_0^4=\rho h\omega^2/D_1=\omega^2/(\tilde{\omega}^2RS). \tag{24}$$

The constants R and S represent the effects of rotatory inertia and transverse shear deformation, respectively. The ratio

$$g=R/S=\pi^2 c_{66}/12(c_{11}-c_{12}^2/c_{22}) \tag{25}$$

plays an important role. It is the ratio of the thickness-shear modulus ($\kappa^2 c_{66}$) to the plate flexure modulus. In the case of an isotropic material, c_{66} becomes the shear modulus, and $c_{11}-c_{12}^2/c_{22}$ reduces to $E/(1-\nu^2)$, where E is Young's modulus and ν is the poisson ratio.

Equations (23) are satisfied by

$$\psi_x=(\sigma-1)\partial\eta/\partial x, \tag{26}$$

where σ is a constant, provided

$$\sigma^{-1}S\delta_0^4=R\delta_0^4-S^{-1}-[S(\sigma-1)]^{-1}=\delta^2, \text{ (say)} \tag{27}$$

and

$$(\partial^2\eta/\partial x^2)+\delta^2\eta=0. \tag{28}$$

Solution of (27) yields

$$2(\delta_1^2, \delta_2^2)=\delta_0^4\{R+S\pm[(R-S)^2+4\delta_0^{-4}]^{\frac{1}{2}}\}, \tag{29}$$

and

$$(\sigma_1, \sigma_2)=(\delta_1^{-2}, \delta_2^{-2})S\delta_0^4. \tag{30}$$

Hence, two deflection functions (η_1 and η_2) are obtained by setting δ equal to δ_1 or δ_2 in (28). They are the solutions of the fourth-order equation

$$[(\partial^2/\partial x^2)+\delta_1^2][(\partial^2/\partial x^2)+\delta_2^2]\eta=0, \tag{31}$$

which can be obtained directly by eliminating ψ_x from (23). Equation (31) is, in fact, no more than an anisotropic plate version of Timoshenko's beam equation.[7]

To sum up, we may write

$$\eta=\eta_1+\eta_2, \tag{32}$$

$$\psi_x=(\sigma_1-1)(\partial\eta_1/\partial x)+(\sigma_2-1)\partial\eta_2/\partial x, \tag{33}$$

$$(\partial^2\eta_1/\partial x^2)+\delta_1^2\eta_1=0, \quad (\partial^2\eta_2/\partial x^2)+\delta_2^2\eta_2=0, \tag{34}$$

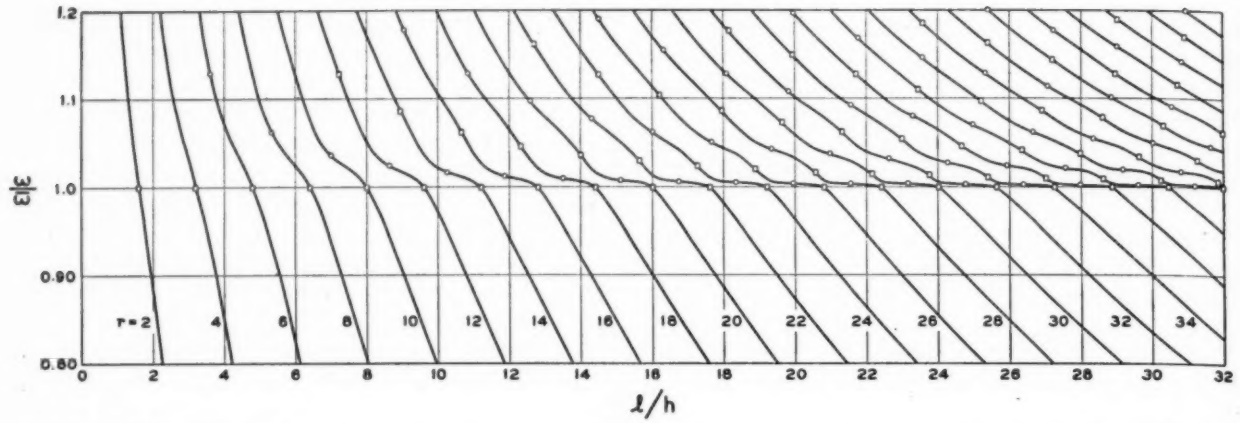


FIG. 1. Resonant frequencies of the even modes of vibration of an AT-cut quartz plate as computed from Eqs. (42), (43), and (44). The ordinate is the ratio of the resonant frequency (ω) to the frequency ($\bar{\omega}$) of the first thickness-shear mode of an infinite plate. The abscissa is the ratio of the length (l) to the thickness (h). The circles and squares locate points for which the transcendental frequency equation can be solved explicitly. They are also the points at which the corresponding curves for the odd modes would cross the curves shown.

and, from (7), $M_{zz}=0$,

$$M_x=-D_1[(\sigma_1-1)\delta_1^2\eta_1+(\sigma_2-1)\delta_2^2\eta_2],$$

and (35)

$$Q_x=D_6[(\sigma_1\partial\eta_1/\partial x)+\sigma_2\partial\eta_2/\partial x].$$

The solutions of (34), which are even in x (i.e., the symmetric, or odd order, flexure modes), make Q_x odd in x. Hence, they cannot annul (22), which is even in x, and play no part in satisfying the boundary conditions on $x=\pm l/2$. This is the counterpart of the experimental observation that there is no coupling between the fundamental thickness-shear mode and the odd flexure modes. The remaining solutions of (34) are,

$$\eta_1=A_1\sin\delta_1 x, \quad \text{and} \quad \eta_2=A_2\sin\delta_2 x. \tag{36}$$

The constants A_1 and A_2 are evaluated from the boundary conditions, i.e., (36) are substituted into (35) and then, at $x=\pm l/2$, $M_x=0$, $Q_x=-D_6B$ to annul (22). This results in the equations

$$A_1(\sigma_1-1)\delta_1^2\sin(\delta_1 l/2)+A_2(\sigma_2-1)\delta_2^2\sin(\delta_2 l/2)=0,$$

and (37)

$$A_1\sigma_1\delta_1\cos(\delta_1 l/2)+A_2\sigma_2\delta_2\cos(\delta_2 l/2)=-B.$$

These equations give A_1 and A_2 in terms of the forcing amplitude B. Resonance occurs when the determinant of (37) vanishes, i.e., when

$$\sigma_2\delta_1(\sigma_1-1)\tan(\delta_1 l/2)=\sigma_1\delta_2(\sigma_2-1)\tan(\delta_2 l/2). \tag{38}$$

The roots of this equation give the resonant frequencies in terms of the dimensions and elastic constants of the plate.

FREQUENCY EQUATION

To put the frequency Eq. (38) into a form suitable for computing the roots, let

$$\gamma=\delta_1 l/2, \tag{39}$$

and

$$a=\delta_2/\delta_1=[(1-b)/(1+b)]^{\frac{1}{2}}, \tag{40}$$

where b is the positive root of

$$b^2=1+4g(\bar{\omega}^2/\omega^2-1)(1+g)^{-2}. \tag{41}$$

It may be seen, from (29), that δ_1 is always real, while δ_2 is real if $\omega>\bar{\omega}$ and imaginary if $\omega<\bar{\omega}$. From (40), the same applies to a. (The case $\delta_2=0$, $a=0$, and $\omega=\bar{\omega}$ will be considered later along with other special roots.) When a is imaginary, let $a=ia_1$, where a_1 is real. Then (38) may be written as

$$(a^2-g)\tan\gamma=a(1-ga^2)\tan a\gamma, \qquad \omega>\bar{\omega},$$

and (42)

$$(a_1^2+g)\tan\gamma=a_1(1+ga_1^2)\tanh a_1\gamma, \qquad \omega<\bar{\omega}.$$

From (40) and (41),

$$\begin{aligned}\omega/\bar{\omega}&=[1-a^2(1+g)^2/g(1+a^2)^2]^{-\frac{1}{2}}, & \omega>\bar{\omega}\\ \omega/\bar{\omega}&=[1+a_1^2(1+g)^2/g(1-a_1^2)^2]^{-\frac{1}{2}}, & \omega<\bar{\omega},\end{aligned} \tag{43}$$

and, from (39),

$$\begin{aligned}l/h&=\gamma(\bar{\omega}/\omega)[(1+a^2)/3(1+g)]^{\frac{1}{2}}, & \omega>\bar{\omega}\\ l/h&=\gamma(\bar{\omega}/\omega)[(1-a_1^2)/3(1+g)]^{\frac{1}{2}}, & \omega<\bar{\omega}.\end{aligned} \tag{44}$$

The elastic constants affect the frequencies only through the constants g and c_{66} (the latter is contained in $\bar{\omega}$). Thus, there is a family of roots of (42) for each value of g, i.e., for each crystal class and orientation of cut. Selection of g and a or a_1 fixes $\omega/\bar{\omega}$ according to (43) and defines roots γ of (42). The values of γ are computed and inserted in (44) to give the values of l/h corresponding to the chosen g and a or a_1, i.e., to the chosen $\omega/\bar{\omega}$. For a given g, this procedure is repeated over the desired range of $\omega/\bar{\omega}$ (or a and a_1).

The results of a set of computations (for $g=0.283$) are shown in Fig. 1. The curves will be recognized as being characteristic of the frequency spectrum of resonances excited by thickness-shear vibration of

rectangular quartz plates.[12] The circled and squared points in Fig. 1 represent special roots of (42) for which both sides of the equation are infinity and zero, respectively. These roots will be discussed in detail later.

COMPARISON WITH FREQUENCIES OF *AT* QUARTZ PLATES

Mason's values[13] for the elastic constants of quartz, referred to the principal axis of crystal symmetry, are, in units of 10^{10} dynes/cm^2, $c_{11}^0=86.05$, $c_{12}^0=5.05$, $c_{13}^0=10.45$, $c_{14}^0=18.25$, $c_{33}^0=107.1$, $c_{44}^0=58.65$, and $c_{66}^0=(c_{11}^0-c_{12}^0)/2=40.5$.

The constants which are required, referred to rotated axes, are given by [14]

$$c_{66}=c_{44}^0s^2+(c_{11}^0-c_{12}^0)c^2/2-2c_{14}^0sc,$$
$$c_{22}=c_{11}^0c^4+c_{33}^0s^4+2(2c_{44}^0+c_{13}^0)s^2c^2+4c_{14}^0sc^3,$$
$$c_{12}=c_{12}^0c^2+c_{13}^0s^2-2c_{14}^0sc, \quad c_{11}=c_{11}^0,$$

where $s=\sin\theta$ and $c=\cos\theta$. For the AT cut, $\theta=35°15'$. Hence, in units of 10^{10} dynes/cm^2, $c_{66}=29.34$, $c_{22}=129.9$, $c_{12}=-10.49$, $c_{11}=86.05$, and, from (25) and (18), $g=0.283$. Thus, the curves in Fig. 1 are for the AT cut. Portions of these curves are drawn to a larger scale in Figs. 2 and 3.

Data obtained by Sykes with AT plates are plotted as circles in Figs. 2 and 3. The dimensions of the plates were initially 25×25×0.807 mm for Fig. 2 and 22×14×1.000 mm for Fig. 3. In successive experiments the x-dimensions (l) were cut down in decrements of 0.1 mm to 21.8 mm and 8 mm, respectively. Hence, l/h varies from 28.38 to 30.98 in Fig. 2 and 8 to 14 in Fig. 3. The recorded resonant frequencies (cps) were converted to $\omega/\bar{\omega}$ by multiplying by $2h/(c_{66}/\rho)^{\frac{1}{2}}=0.485\times10^{-6}$, in which $h=0.807$ and 1.000, $\rho=2.65$, and $c_{66}=29.34\times10^{-10}$. A better fit between experiment and theory would have been obtained in the neighborhood of $\omega/\bar{\omega}=1$ if c_{66} had been calculated from Sykes' data on the AT plate rather than from Mason's primary data on the nonrotated constants.

It may be seen that there is a relative displacement, between theory and experiment, of about 0.1 to 0.3 percent along the l/h axis. This could be due to approximations made in the theory or to errors in the elastic constants. In the latter case, for example, a 1 percent error in c_{11} would produce a 0.1 percent displacement of the type observed.

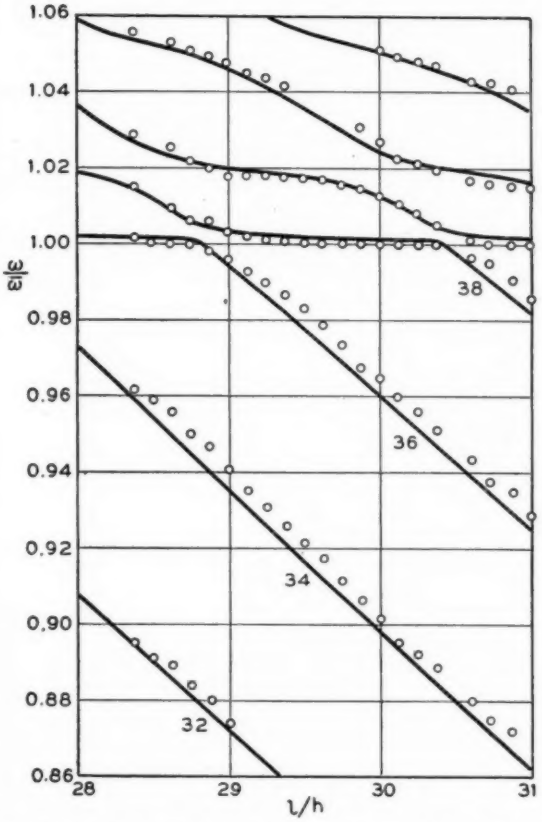


FIG. 2. Comparison of theory with data obtained by Sykes on an AT quartz plate, initially 25×25×0.807 mm, cut down in decrements of 0.1 mm to 25×21.8×0.807 mm.

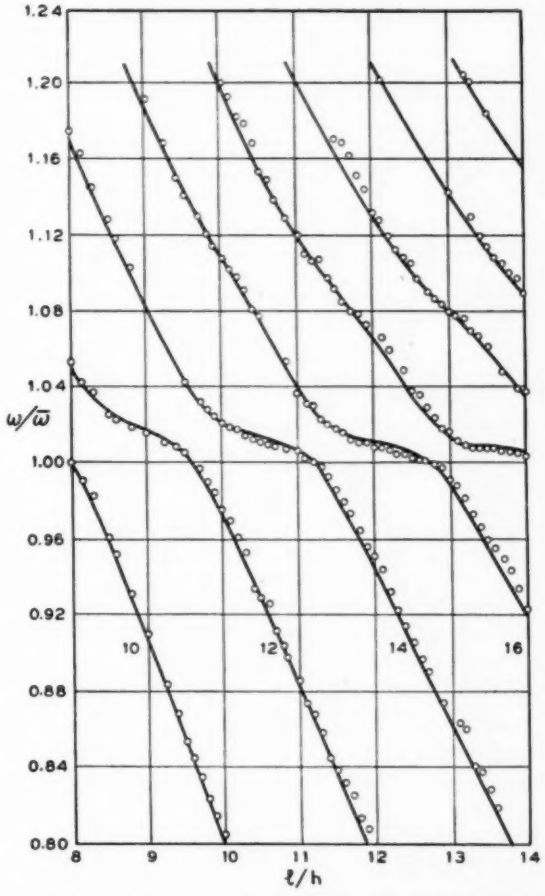


FIG. 3. Comparison of theory with data obtained by Sykes on an AT quartz plate, initially 22×14×1.000 mm cut down in decrements of 0.1 mm to 22×8×1.000 mm.

[12] See reference 3, p. 233, Fig. 6.15.

[13] W. P. Mason, *Piezoelectric Crystals and Their Applications to Ultrasonics* (D. Van Nostrand Company, Inc., New York, 1950), p. 84.

[14] See reference 3, p. 247.

SOLUTION OF FREQUENCY EQUATION BY SUCCESSIVE APPROXIMATIONS

Special interest centers around the frequencies of the higher modes in the neighborhood of the thickness-shear frequency $\bar{\omega}$. In this range, the following solution of the frequency equation by successive approximations (suggested by S. A. Schelkunoff) converges very rapidly. Writing the first of (42) in the form

$$\tan\gamma=\varphi(g,a)\tan a\gamma, \quad \omega>\bar{\omega}, \tag{45}$$

let

$$\gamma=r\pi+\beta, \quad -\pi/2<\beta<\pi/2, \quad r=1,2,3\ldots.$$

Since $\tan(r\pi+\beta)=\tan\beta$, (45) may be written

$$\beta=\tan^{-1}[\varphi(g,a)\tan a(r\pi+\beta)], \quad \omega>\bar{\omega}.$$

As a first approximation take

$$\beta_1=\tan^{-1}[\varphi(g,a)\tan ar\pi], \quad \omega>\bar{\omega},$$

and, as successive approximations,

$$\beta_{i+1}=\tan^{-1}[\varphi(g,a)\tan a(r\pi+\beta_i)], \quad \omega>\bar{\omega}.$$

It is usually necessary to take only one or two steps.

Similarly, for $\omega<\bar{\omega}$,

$$\tan\gamma=\varphi_1(g,a_1)\tanh a_1\gamma, \quad \omega<\bar{\omega}, \tag{46}$$

and

$$\beta_{i+1}=\tan^{-1}[\varphi_1(g,a_1)\tanh a_1(r\pi+\beta_i)], \quad \omega<\bar{\omega}.$$

The product $a_1 r\pi$ does not have to become very large before this can be written as

$$\beta=\tan^{-1}\varphi_1(g,a_1),$$

which accounts for the monotonic, regularly spaced curves for the low frequency portions of the flexural modes in Fig. 1.

ODD MODES

While only the even flexural modes are excited by the forcing function (19), the odd modes may be excited by forcing a thickness-shear deformation antisymmetrical with respect to x. In this case, the residual stress resultants corresponding to (22) will be odd in x and the even solutions

$$\eta_1=B_1\cos\delta_1 x, \quad \eta_2=B_2\cos\delta_2 x \tag{47}$$

of (34) will be required to annul them. The only change in the results, therefore, is the interchange of sine and cosine throughout. The transcendental equations are the same as those for the even modes except for the replacement of $\varphi(g,a)$ and $\varphi_1(g,a_1)$ by their reciprocals and a reversal of sign for the case $\omega/\bar{\omega}<1$. The treatment of these equations is then very similar. In particular, the circled and squared points are also roots and hence the odd-mode curves of $\omega/\bar{\omega}$ *vs* l/h, corresponding to Fig. 1, cross the curves of Fig. 1 at these points. Whereas, for the even modes, the circled points are on the flat portions of the curves (slow variation of $\omega/\bar{\omega}$ with l/h) and the squared points are on the steep portions (rapid variation of $\omega/\bar{\omega}$ with l/h), the converse is true for the odd flexural modes.

ALGEBRAIC FORMULAS

The circled points in Fig. 1 are those for which γ and $a\gamma$ in the first of (42) are odd multiples of $\pi/2$. In the case of excitation by thickness-shear, even in x, these points are of special interest, as they are far from the frequencies of neighboring modes and center in regions where the frequency changes slowly with change of dimensions. The squared points, for which γ and $a\gamma$ in the first of (42) are even multiples of $\pi/2$, have the same properties for thickness-shear excitation odd in x. Because of the degeneracy of the transcendental equation for both sets of points, algebraic formulas may be obtained for them.

In the first of (42), the roots under discussion are

$$\begin{aligned}\gamma&=r\pi/2, \quad r=1,2,3\ldots\\ a\gamma&=n\pi/2, \quad n=0,1,2\ldots,\end{aligned} \tag{48}$$

from which

$$a=n/r. \tag{49}$$

Substituting (48) and (49) in the first of (43) and (44), the latter become

$$\omega/\bar{\omega}=g^{\frac{1}{2}}(r^2+n^2)[(gr^2-n^2)(r^2-gn^2)]^{-\frac{1}{2}} \tag{50}$$

$$l/h=(\pi/2)[(gr^2-n^2)(r^2-gn^2)/3g(1+g)(r^2+n^2)]^{\frac{1}{2}}. \tag{51}$$

Equations (50) and (51) give the coordinates ($\omega/\bar{\omega}$, l/h) of the circled points (odd r and n) and squared points (even r and n). These points are plotted in Fig. 4 for the AT plate ($g=0.283$) over the ranges 0 to 1.2 for $\omega/\bar{\omega}$ and 0 to 32 for l/h.

The equations of the two families of curves passing through the points $n=$ const and $r=$ const are obtained by eliminating r and n, respectively, between (50) and (51) with the results

$$2(\omega/\bar{\omega})^2=1+(kn^2h^2/l^2)+[(1+kn^2h^2/l^2)^2-4(\kappa nh/l)^4/g]^{\frac{1}{2}}, \tag{52}$$

and

$$2(\omega/\bar{\omega})^2=1+(kr^2h^2/l^2)-[(1+kr^2h^2/l^2)^2-4(\kappa rh/l)^4/g]^{\frac{1}{2}}, \tag{53}$$

where

$$k=\kappa^2(1+g)/g=(\pi^2/12)+(c_{11}c_{22}-c_{12}^2)/c_{22}c_{66}. \tag{54}$$

These curves are illustrated in Fig. 4 for the AT cut.

Equation (52) is to be compared with the empirical equation given by Sykes which, for the present case and in the present notation, is

$$(\omega/\bar{\omega})^2=1+kn^2h^2/l^2, \tag{55}$$

where k is an experimental constant. Sykes' equation was derived by Ekstein, from his theory, with a formula for k more elaborate than (54) but yielding approximately the same numerical values. Equations (55) and (52) differ, in form, only by the term $4(\kappa nh/l)^4/g$ in the

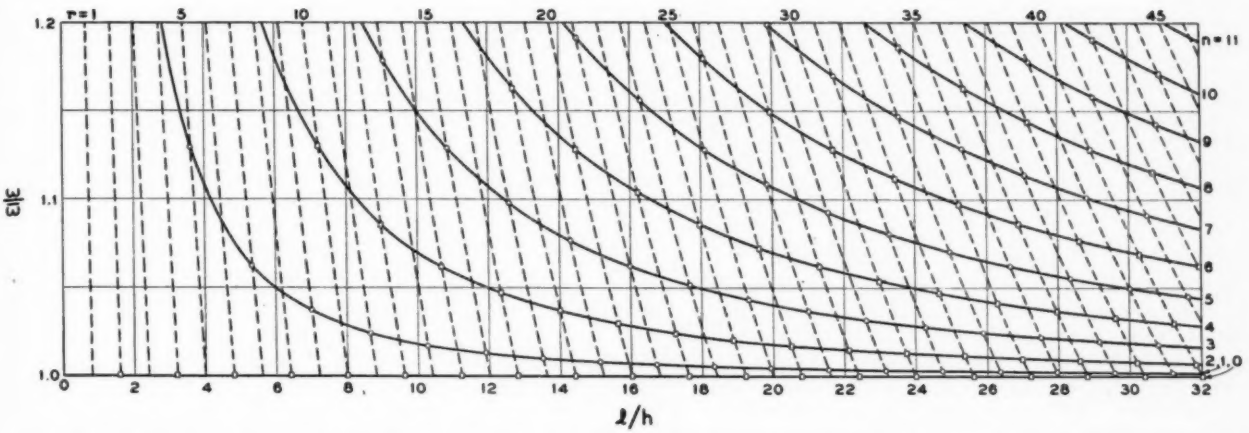


FIG. 4. The circles and squares are those from the upper half of Fig. 1. Their coordinates are given by the algebraic formulas (50) and (51). They are connected by full and dashed curves given by Eqs. (52) and (53), respectively. Points on these curves do not correspond to resonant frequencies except at the intersections marked with a circle or square. The full curves $n=2, 3, 4 \ldots$ are those commonly designated as the first, second, third . . . overtones of the fundamental thickness-shear mode.

latter. The additional term accommodates the higher overtones of thickness-shear vibration (large n) and small values of l/h.

It should be observed, however, that equations like (52) or (55) have meaning only at discrete points. These may be located by (50) and (51) or by the intersections of the families of curves (52) and (53) as illustrated in Fig. 4.

Regarding Eq. (53), it may be observed that the curves $r=$constant are continuations, into the region $\omega/\bar{\omega}>1$, of the flexural mode curves (Fig. 1) from the region $\omega/\bar{\omega}<1$. Alternatively, curves $n+r=$constant are close approximations to the curves in the upper region of Fig. 1.

SHAPES OF MODES

The shapes of the modes of vibration may be determined very easily for the frequencies discussed in the preceding section. Noting that $\delta_1=r\pi/l$ and $\delta_2=n\pi/l$, we find, from (32), (33), and (36),

$$\eta=A_1\sin(r\pi x/l)+A_2\sin(n\pi x/l), \tag{56}$$

and

$$h\psi_x=U_1\sin(r\pi x/l)+U_2\sin(n\pi x/l), \tag{57}$$

where

$$U_1=A_1h(\sigma_1-1)\delta_1, \quad U_2=A_2h(\sigma_2-1)\delta_2.$$

These give, respectively, the displacements that would be observed normal and tangential to the surface of the plate. Each contains a short wavelength part, $\sin(r\pi x/l)$, and a long wavelength part, $\sin(n\pi x/l)$. The former is commonly associated with the notion of flexural modes and the latter with thickness-shear modes. Alternatively, each type of mode contains both normal and tangential components of displacement.

The relative amplitudes of the four parts of the surface displacement may be found by using the first of (37), from which

$$\left|\frac{A_2}{A_1}\right|=\left|\frac{(\sigma_1-1)\delta_1^2\sin(\delta_1 l/2)}{(\sigma_2-1)\delta_2^2\sin(\delta_2 l/2)}\right|. \tag{58}$$

Noting that $\sigma_1/\sigma_2=a^2=n^2/r^2$ and $\sigma_1=(1+a^2)/(1+g)$, this becomes, for r and n odd,

$$|A_2/A_1|=(gr^2-n^2)/(r^2-gn^2). \tag{59}$$

For r and n even, the even solutions (47) are used, and the resulting amplitude ratio is the same as (59). Usually r/n is large, and hence $|A_2/A_1|$ is approximately equal to g. Thus, regarding displacements normal to the plate, the part associated with thickness-shear is less than that associated with flexure in the ratio $g:1$ (approximately). The corresponding ratio for the tangential displacements is

$$U_2/U_1=r/n;$$

i.e., the tangential displacement associated with thickness-shear is usually much larger than that associated with the accompanying flexural motion. Similarly, it can be shown that $|U_2/A_2|$ is approximately r/ng times $|U_1/A_1|$; also, $|U_2/A_1|$ is r/n times $|U_1/A_1|$. Thus, in general, for the circled points the predominant motion is tangential to the surface of the plate.

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

A Correlating Modulus for Fluid Resistance in Accelerated Motion*

H. W. IVERSEN
University of California, Berkeley, California

AND

R. BALENT
University of Denver, Denver, Colorado

(Received August 28, 1950)

The resistance to accelerating movement of objects in a fluid is normally evaluated with a mass term in the equation of motion that is greater than the actual mass of the object by an added mass described as a constant times the displaced mass of fluid. This added mass has been evaluated by numerous investigators for different body shapes from potential flow considerations. Published experimental results, mostly with oscillating systems with small amplitudes of motion, show an added mass constant that is higher than that derived from potential flow with values that are dependent upon the fluid and the object size. A few previous experiments on resistance in unidirectional accelerated motion indicate that the added mass is variable and depends upon the state of motion.

The present investigation includes a development from model law considerations that results in a resistance equation of the form

$$F=C\tfrac{1}{2}\rho V^2S.$$

This is identical with the normal resistance equation for steady state except for the coefficient C which, in addition to being a function of Reynolds' modulus, Froude's modulus, and the geometry, is also a function of a modulus, AL/V^2, where A is the acceleration, L is a characteristic length, and V is the velocity.

Experiments with circular disks moving perpendicular to the plane of the disk under the action of approximate constant driving forces show this resistance coefficient to be correlated by the modulus AL/V^2.

INTRODUCTION

THE resistance of objects in accelerated motion in a fluid medium has been investigated both analytically and experimentally. DuBuat,[1] in 1786, found by experiment that the motion of spheres oscillating in water could be described only if an added mass which was greater than the actual mass of the sphere was included in the equation of motion. Bessel,[2] in 1820, represented this added mass by a factor, k, defined as

$$M_a=(k)\text{ (mass of fluid displaced by the body)}, \quad (1)$$

where M_a = added mass. Values of k obtained by numerous investigators are listed in Table I.

These references are summarized to show the methods of attack and the variations of the results of the k factors. The general conclusions that can be drawn from the summary are that the added mass (or "Virtual Mass," as it is sometimes designated in the literature) is a function of

(1) the size and shape of the body,
(2) the direction in which it is moved through the fluid with respect to an axis in the body, and
(3) the density and viscosity of the fluid.

The equation of unidirectional accelerated motion in which these k factors are used is written

$$F-M_eA=C_D\tfrac{1}{2}\rho V^2S+k\rho BA, \quad (2)$$

where F = net driving force, M_e = mass of the solid, A = acceleration, C_D = drag coefficient, ρ = fluid density, V = velocity, S = area, and B = displaced volume of fluid. The drag coefficient is taken as that which corresponds to the steady state velocity of the value V. However, some experimental results showed that this added mass effect could not be described by a constant value of k, but depended upon the state of the motion. In some cases empirical curve fitting was applied to describe the motion.[3]

An examination of the equation of motion yields a development which correlates experimental data (see Appendix). The resistance equation for objects moving in accelerated unidirectional motion is shown to be of the same form as that used for steady state motion

$$\bar{F}=C\tfrac{1}{2}\rho V^2S, \quad (3)$$

where $\bar{F}$ = net fluid resistance. In the case of accelerated motion

$$C=\varphi(N_R, N_F, \text{geometry}, AL/V^2), \quad (4)$$

where N_R = Reynolds' modulus, N_F = Froude's modulus, and L = characteristic length of the body. This modulus, AL/V^2, which includes both acceleration and velocity, should therefore be a correlating modulus for accelerated flow.

EXPERIMENT

The validity of Eq. (3) was checked by experiment. In order to minimize the effect of viscous and gravity force fields, the object used was a flat disk which was submerged in water and moved perpendicular to the plane of the disk. The disk was suspended by a triangular wire harness with a flexible wire cable from the harness over an aluminum pulley to weights which provided the driving force (Fig. 1). External friction was

* R. Balent, "Virtual mass of disks," Master's thesis, University of California, 1949. (The present paper is a summary of this work.)

[1] C. DuBuat, *Principes d'hydraulique* (Paris, 1786).

[2] F. W. Bessel, *On the Incorrectness of the Reduction to a Vacuum Formerly Used in Pendulum Experiments* (Berlin Academy, 1826).

[3] R. G. Lunnon, Proc. Roy. Soc. (London) **A118**, 680–694 (1928).

TABLE I. Summary of investigations on added mass.

Investigator	Shape of object	Type of motion	Fluid	Added mass factor k	Remarks
Bessel[a]	Sphere	Oscillating	Air Water	0.9 0.6	Experiments with spherical pendulum
Poisson[b]	Sphere	Unidirectional	Ideal	0.5	Analytical evaluation from potential flow
Stokes[c]	Sphere	Unidirectional	Viscous	0.5+	Analytical evaluation with laminar flow
McEwen[d]	Sphere	Oscillating	Viscous	0.5+	Experimental
Krishnaiyer[e]	Sphere	Oscillating	Water	0.58	Experimental
Smith[f]	Elongated body of revolution	Unidirectional	Air	0.072	Study of airship *Akron*
Taylor[g]	Elongated body of revolution	Linear	Ideal	0.051	Airship hull shape from potential flow
Browne, Moullin, and Perkins[h]	Rectangular bars	Vertical oscillation	Water	0.65	Experimental with floating prisms
Murphy[i]	Ellipsoids and ship models	Unidirectional	Water	0.03 to 0.07	Referred to displaced mass of water
Lamb[j]	Disk	Perpendicular to plane of disk	Ideal	0.637	Referred to volume of a sphere with diam. equal to disk diam.
Yee-Tak-Yu[k]	Disk	Oscillating	Water, gasoline, and carbon tetrachloride	0.81	Experimental
Gracey[l]	Flat plates	Oscillating	Air	0.94 and 0.96	Pendulum oscillating with different length, breadth ratio
Lunnon[m]	Spheres	Unidirectional	Water	0.5 to 2.0	Experimental

[a] See reference 2.
[b] S. D. Poisson, Men. de l'institute **11**, 521–582 (1832).
[c] Sir G. G. Stokes, Trans. Cambridge Phil. Soc. **9**, 8–106 (1850).
[d] G. F. McEwen, Phys. Rev. **33**, 429–511 (1911).
[e] N. C. Krishnaiyar, Phil. Mag. **6**, No. 46, 1049–1053 (1923).
[f] R. H. Smith, J. Aeronaut. Sci. **3**, No. 1, 26–31 (September, 1935).
[g] G. I. Taylor, Proc. Roy. Soc. (London) **A120**, 13–21 (1928).
[h] Browne, Moullin, and Perkins, Proc. Cambridge Phil. Soc. **26**, 258–272 (1930).
[i] J. P. Murphy, "Determination and correlation of the virtual mass of ship models," unpublished Master's thesis, University of California, 1937.
[j] H. Lamb, *Hydrodynamics* (Cambridge University Press, London, 1924), sixth edition, pp. 138–139, 144.
[k] Yee-Tak-Yu, J. Appl. Phys. **13**, 66–69 (1942).
[l] W. Gracey, "Additional mass effect of plates as determined by experiments," NACA Report No. 707.
[m] See reference 3.

reduced to a minimum with ball bearings on the pulley shaft. A graduated scale on the periphery of the pulley together with a stationary pointer was photographed with a 16-mm movie camera to indicate the pulley position and hence the disk position. A sweep second synchronous clock was also placed in the field of view of the camera to give accurate timing. The disk was submerged in a tank 8 ft×8 ft×20 ft deep with its initial position approximately four feet from the bottom. Upon release of the cable by a solenoid trip the disk rose in the tank under the action of the falling weights.

Data were taken with two different size disks, 2 feet and 1 foot in diameter. Four different driving forces were applied to each disk. Table II lists the test conditions. The initial driving force was measured by a beam balance under the driving weight with the system free and the disk submerged at the starting position. Included in Table II are the actual masses of each moving system. A correction to the driving force was necessary because of the transfer of the cable from one side of the system to the other. This can be expressed as

$$F_x = F_1 + 0.07X, \tag{5}$$

where x is the distance of travel from the initial starting position and F_1 is the initial force.

Results were interpreted from the distance-time relationships which were obtained from the pulley position and the time as recorded on the frames of the movie

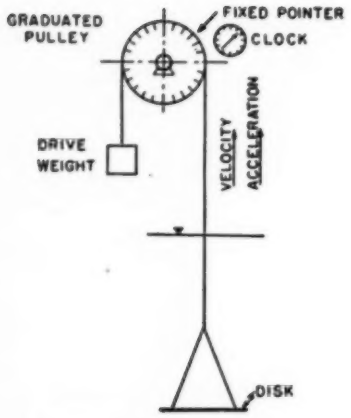


FIG. 1. Diagram of apparatus.

strips. Velocities and accelerations were obtained by graphical differentiation of the distance-time curves. These are shown in Fig. 2.

All necessary data for evaluation of the coefficient C from Eq. (3) and the modulus AL/V^2 were then available. For this experimental system the net fluid resistance is

$$\bar{F}=F_x-M_eA. \tag{6}$$

Results are shown on Fig. 3 with the disk diameter D as the characteristic length. The points shown on Fig. 3 were evaluated from the curves of Fig. 2 at each 0.2 second up to one second and for each half-second thereafter to four seconds for the one-foot disk, and at each half-second up to three seconds and at each one second thereafter to eight seconds for the two-foot disk.

DISCUSSION

The agreement of the results, when correlated in this manner, show the validity of the correlating modulus of Eq. (4). The Reynolds modulus does not have any influence on the resistance coefficient for the system of a disk moving perpendicular to the plane of the disk except at the onset of motion. The steady-state drag coefficient of 1.12 is reported[4] to hold for Reynolds numbers above 10^3. The corresponding values of the velocities of the disks of this experiment were 0.005 feet per second for the two-foot disk and 0.01 feet per second for the one-foot disk. Reference to Fig. 2 shows the corresponding largest time interval from the start of motion to be insignificant. Likewise, since the disks were completely submerged at all times and did not produce any surface disturbances in the fluid, the Froude modulus does not influence the resistance coefficient.

When the acceleration is zero, the modulus AL/V^2 is zero, the resistance coefficient should be that of the steady-state drag coefficient. Results from the one-foot disk, Fig. 3, show this approach to the value 1.12. The two-foot disk results indicate an approach to a lower value of the steady-state drag coefficient. However, side-wall interference effects become apparent for this size of object in the tank which was used.[4] The trend of the results at terminal velocities from this effect would be as shown. A curve has been drawn through the plotted results which follows the results of the one-foot disk at the smaller values of the modulus.

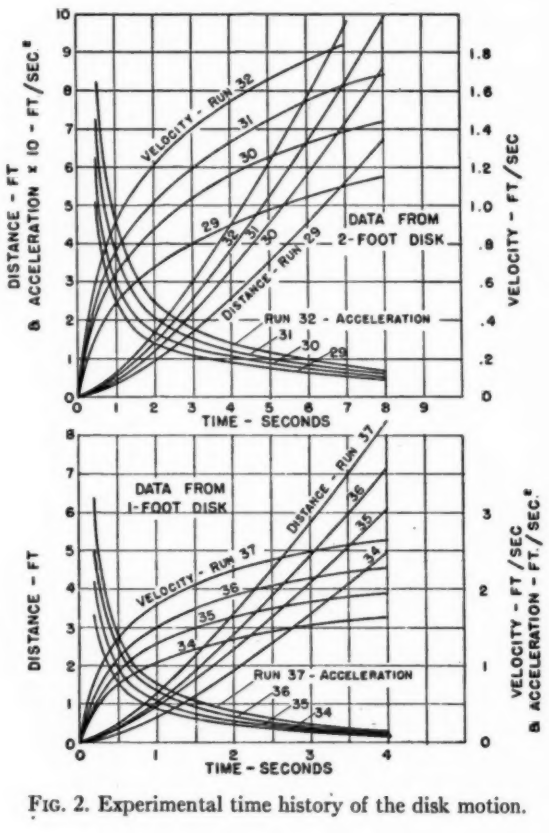


FIG. 2. Experimental time history of the disk motion.

TABLE II. Test conditions.

Run	Disk diameter (feet)	Gross driving force (pounds)	Mass of moving parts (slugs)
29	2	4.78	1.263
30	2	6.78	1.325
31	2	8.78	1.387
32	2	10.78	1.449
34	1	2.28	0.512
35	1	3.28	0.543
36	1	4.28	0.574
37	1	5.28	0.605

Values of the added mass factor, k, from Eq. (2) were evaluated from the data with the drag coefficient, C_D, taken as 1.12. The displaced mass of fluid was taken as that of a sphere with a diameter equal to the disk diameter. Added mass factors are shown in Fig. 4 as a function of the modulus AD/V^2. A correlation results except at the low values of the modulus. Discrepancies in this range are due to the aforementioned side wall interference effects and also to the comparative accuracy in this range. The left side of Eq. (2) and the first

[4] Dryden, Murnaghan, and Bateman, "Hydrodynamics," Bull. Nat. Research Council, No. 84, 14–15 (1932).

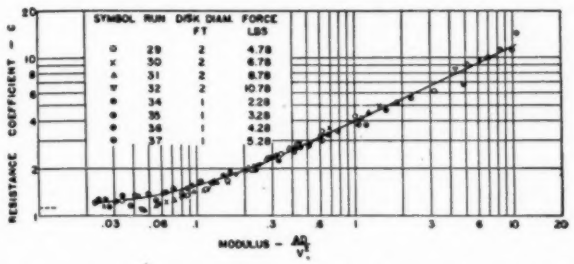


FIG. 3. Resistance coefficient for disks in accelerated motion in a fluid under the action of a constant driving force.

term on the right become of the same order of magnitude. The difference of these two terms becomes of doubtful accuracy, which is directly reflected in the value of k. The curve which is shown was computed from the curve of Fig. 3.

When the velocity is small, i.e., at the start of the motion, the modulus AD/V^2 is large. The value of k approaches that of 0.637 developed from potential flow considerations. As the velocity becomes of appreciable value, the added mass factor increases to approximately five times the minimum value. Hence, use of a constant value of the added mass factor with drag coefficients which correspond to steady-state values of the velocities to evaluate accelerated motion may be considerably in error.

A qualitative description of the trends of Fig. 4 may be made from a consideration of the flow patterns about the disk as diagrammed in Fig. 5. At the onset of motion of the present experimental system, the fluid motion corresponds to that of potential flow (Fig. 5a). As the velocity becomes significant, the wake is formed which consists of eddies on the downstream side of the disk (Fig. 5b). At high values of the modulus AD/V^2 the flow corresponds to Fig. 5a and the added mass factor would be expected to be the value computed from potential flow consideration. At low values of the modulus AD/V^2 there is a mass in the wake on the downstream side of the disk that must be accelerated with the disk. This mass is larger than the disturbed mass of Fig. 5a, and hence, k should be larger under these conditions.

The distance-time history of a disk under a known driving force may then be evaluated from Fig. 3 or Fig. 4 and Eq. (3). Since the resistance coefficient is dependent upon both the velocity and acceleration, a graphical integration process is indicated to obtain distance and time. No attempt has been made to express the resistance coefficient as an algebraic function of the correlating modulus. One method of solution is as follows.

(1) With the force, area, and fluid known, select values of velocity less than the terminal or steady-state velocity for the applied force.

(2) Compute the resistance coefficient from Eq. (3).

(3) Obtain the modulus AD/V^2 from Fig. (3).

(4) Compute the corresponding acceleration. If the mass of the moving parts is appreciable, a correction can be made to include the acceleration of this mass.

(5) Obtain time as a function of velocity from a graphical integration of $t=\int_0^V dV/A$.

(6) Obtain distance as a function of velocity from a graphical integration of $x=\int_0^t V dt$ or from a graphical integration of $x=\int_0^V V dV/A$.

At low velocities Fig. 4 and Eq. (2) can be used to advantage to describe the curves of step (5) and (6) at the times and velocities approaching zero.

A number of further considerations are indicated from the results of this use of the correlating modulus as described.

1. Does the resistance coefficient of Fig. 3 apply to decelerating motion of disks?

2. Will the resistance coefficient of Fig. 3 apply if the initial velocity is not zero, but is some finite steady state value?

3. Will this type of correlation hold for other objects such as spheres, and what are the resistance coefficients for other objects?

4. How can the Reynolds number in the case of viscous effects, or the Froude number in the case of surface effects, be incorporated into the correlation?

Further work is planned to include some of these conditions.

CONCLUSIONS

The results of the experiments and analysis presented in this discussion show

(1) a resistance coefficient for accelerated motion which is a function of the modulus, AD/V^2 with the equation of motion stated as,

$$\bar{F}=C\tfrac{1}{2}\rho V^2 S, \text{ and}$$

(2) values of C for disks moving perpendicular to the plane of the disk in unidirectional motion starting from rest.

APPENDIX

The fluid which is in the field of disturbance of an object moving through the fluid flows around the object. When the relative velocity is steady, i.e., no acceleration of the body relative to the undisturbed fluid, the normal evaluation of the force existing on the body is by a drag coefficient,

$$C_D=F/\tfrac{1}{2}\rho V^2 S, \tag{7}$$

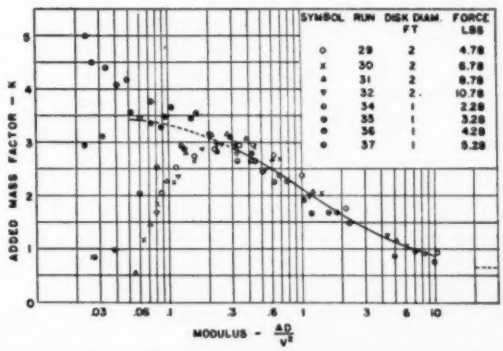


FIG. 4. Added mass factor for disks in accelerated motion in a fluid under the action of a constant driving force.

FIG. 5. Flow lines about a disk: (A) at the onset of motion, (B) with an established wake.

where C_D = drag coefficient = φ (N_R, N_F, geometry), F = force, ρ = fluid density, V = velocity, S = area, N_R = Reynolds modulus, and N_F = Froude modulus. The drag coefficients are usually determined by experiment.

The addition of an acceleration to the motion produces an added resistance which can also be developed in terms of a resistance coefficient and a correlating modulus from a consideration of the various terms of the Navier-Stokes equations. The Navier-Stokes equation written for one axis of an incompressible fluid particle is,

$$\rho Du/Dt = \rho X - (\partial p/\partial x) + \mu \nabla^2 u, \tag{8}$$

where u = particle velocity in x direction,

$Du/Dt = (\partial u/\partial t) + u(\partial u/\partial x) + v(\partial u/\partial y) + w(\partial u/\partial z)$,
x, y, z = perpendicular coordinate directions,
v, w = velocities in y and z directions,
X = body force per unit mass,
p = pressure,
μ = coefficient of absolute viscosity,
∇^2 = Laplace operator = $(\partial^2/\partial x^2) + (\partial^2/\partial y^2) + (\partial^2/\partial z^2)$.

The criteria for dynamical similarity may be developed from this equation. For two systems which are geometrically and dynamically similar, the ratios of the variables are:

Variable	Ratio	Variable	Ratio	(9)
Length	$b_l = l_1/l_2$	Density	$b_\rho = \rho_1/\rho_2$	
Time	$b_t = t_1/t_2$	Velocity	$b_v = v_1/v_2$	
Pressure	$b_p = p_1/p_2$	Acceleration	$b_a = a_1/a_2$	
Viscosity	$b_\mu = \mu_1/\mu_2$	Body force	$b_X = X_1/X_2$.	

The subscripts 1 and 2 refer to the two similar systems under consideration.

Equation (8) written with subscripts 1, designates the flow in system one. Substitution of the ratios of Eqs. (9) give the equation for the second system. Since this dynamical system is one which has changes in velocity with respect to time, the term $\partial u/\partial t$ will be designated as an acceleration, a.

$$b_\rho b_a \rho_2 a_2 + b_\rho (b_v^2/b_l)[\rho_2 u_2(\partial u_2/\partial x_2) + \cdots] = b_\rho b_X \rho_2 X_2 - (b_p/b_l)(\partial p_2/\partial x_2) + b_\mu (b_v/b_l^2)(\mu_2 \nabla^2 u_2). \tag{10}$$

This expression must be the same, for dynamical and geometrical similarity, as Eq. (8) written directly for the second system.

$$\rho_2 a_2 + \rho_2[u_2(\partial u_2/\partial x_2) + \cdots] = \rho_2 X_2 - (\partial p_2/\partial x_2) + \mu_2 \nabla^2 u_2. \tag{11}$$

Comparison of Eqs. (10) and (11) shows,

$$\underset{\text{I}}{b_\rho b_a} = \underset{\text{II}}{b_\rho (b_v^2/b_l)} = \underset{\text{III}}{b_\rho b_X} = \underset{\text{IV}}{b_p/b_l} = \underset{\text{V}}{b_\mu (b_v/b_l^2)}. \tag{12}$$

Each of the terms of Eq. (12) represent force ratios which can be designated as those due to (I) local inertia, (II) convective inertia, (III) gravity, (IV) pressure, and (V) viscosity. For systems where the gravity and viscous fields are negligible, only I, II, and IV need be considered. The pressures are due to the object influence. An integration of pressures on the body will result in the resistance to motion of the body. Hence, the local and convective inertias can be used to define the conditions under which the pressure forces are similar.

Thus,

$$b_\rho b_a = b_\rho (b_v^2/b_l) = b_p/b_l, \tag{13}$$

$$b_a b_l/b_v^2 = 1 = b_p/(b_l b_\rho b_v^2/b_l). \tag{14}$$

In the application of these ratios to the two dynamical systems, any corresponding velocity or acceleration which defines the motion may be taken to evaluate the ratio between the systems. Thus, the free stream velocity and acceleration, or, in the case of an object moving through a stationary fluid, the velocity and acceleration of the object relative to the fluid at rest define the motion.

Hence,

$$\frac{p_2}{\rho_2 V_2^2} = \frac{p_1}{\rho_1 V_1^2}\left(\frac{A_2 L_2}{V_2^2} \Big/ \frac{A_1 L_1}{V_1^2}\right); \tag{15}$$

also,

$$F = \int_0^S p\,dS \propto pL^2.$$

where F is the force on the object and p is the pressure at the boundary of the object of area S.

$$\frac{F_2}{\rho_2 L_2^2 V_2^2} = \frac{F_1}{\rho_1 L_1^2 V_1^2}\left(\frac{A_2 L_2}{V_2^2} \Big/ \frac{A_1 L_1}{V_1^2}\right). \tag{16}$$

When geometrical and dynamical similarity exists, the ratio

$$\frac{A_2 L_2}{V_2^2} \Big/ \frac{A_1 L_1}{V_1^2} = 1,$$

and measurements of forces in one system may be used to predict forces in the second system. Then,

$$F/\rho L^2 V^2 = \varphi(AL/V^2) = C. \tag{17}$$

Equation (17) under the conditions previously stated thus gives the correlating function for the resistance coefficient under accelerated motion. If the viscous and gravity forces are not negligible, a similar analysis shows,

$$C = \varphi(AL/V^2, VL\rho/\mu, V^2/gL). \tag{18}$$

In any case, similar to the concept of the drag coefficient for steady-state conditions, C is also a function of the shape and orientation of the body.

Cut-Off Frequency in Two-Dielectric Layered Rectangular Wave Guides

J. Van Bladel* and Thomas J. Higgins†
University of Wisconsin, Madison, Wisconsin
(Received August 31, 1950)

Equations for the modes and eigenvalues of two-dielectric layered rectangular wave guides with cross sections as in Fig. 1(a), (b), and (c) are derived. From these equations are plotted graphs of cut-off frequency over a range of geometric and dielectric parameters sufficiently wide to cover most requirements of design.

IF a rectangular wave guide is partially filled with a solid dielectric arranged as indicated in Figs. 1(a), (b), and (c), the introduction of this dielectric (a) results in lowering the cut-off frequency and phase velocities of the modes as compared with that of a guide of the same dimensions wholly filled with air, and (b) for a given cut-off frequency enables use of guides of smaller cross section. These effects have led to several important technical uses of layered-dielectric rectangular guides: for example, as a component of airborne antennas.

Little has been written on dielectric layered rectangular guides; what has been published is commented upon in the Appendix. The prime purpose of this paper is (a) to advance a correct and complete analysis of the normal modes for the three configurations of Figs. 1(a), (b), and (c), and (b) to advance graphs which enable determination of the cut-off frequency for each of these configurations. This cut-off frequency will be expressed as a function of the geometric and dielectric parameters of the cross section. Inasmuch as main attention centers on these graphs, the details of analysis underlying their construction have been made as concise as feasible.‡

The usual assumptions underlie the analysis: namely, that the metallic walls are perfectly conducting, the dielectrics are lossless, and the phenomena are harmonic in time. It is not necessary to restrict one dielectric to be air. In general, the two media can be any two dielectrics: we distinguish between the two by terming that with the larger product $\epsilon\mu$ the "hard" medium, and the other the "soft" medium. The hard medium is indicated by the shaded part of the cross section. It is convenient to express the frequency in terms of the wavelength λ_s of a plane wave propagated with the same frequency in an infinitely extending soft medium.

In general, the electromagnetic field in the guide is a linear combination of elemental waves, infinite in number, and termed normal modes. These modes are to be found by taking a z (axial) dependence of the form e^{jhz} for the six components of $\bar{E}$ and $\bar{H}$ and by constructing, through separation of variables, a wave which matches all boundary conditions. This procedure, as is illustrated below, yields a relation between h^2 and ω termed the characteristic equation, of such form that for a given angular frequency ω, h^2 can take only a discrete, though infinite, set of values (the eigenvalues). To each eigenvalue corresponds a mode. When the eigenvalue h^2 is negative, h is imaginary, e^{jhz} is a damping factor and the mode is not propagated. When h^2 is positive, the mode is propagated. The transition between these two states occurs at a frequency, termed the cut-off frequency of the mode, for which h^2 is equal to zero. The lowest of these modal cut-off frequencies is the general cut-off frequency of the guide, below which no energy can be propagated in the guide. We proceed to analytic determination of these desired frequencies.

It follows from Maxwell's equations that the field components of a mode can be expressed as a function of two of them, E_z and H_z, by the equations

$$\{(\omega^2/c^2)-h^2\}[\bar{E}_c]=k^2[\bar{E}_c]=jh\,\mathrm{grad}[E_z]+j\omega\mu(\bar{K}x\,\mathrm{grad}[H_z]), \quad (1)$$

$$\{(\omega^2/c^2)-h^2\}[\bar{H}_c]=k^2[\bar{H}_c]=jh\,\mathrm{grad}[H_z]-j\omega\epsilon(\bar{K}x\,\mathrm{grad}[E_z]), \quad (2)$$

where the subscript c denotes a component in the phase of the cross section, $\bar{K}$ is the unit-vector directed along the z-axis, and the square brackets denote the (x, y)

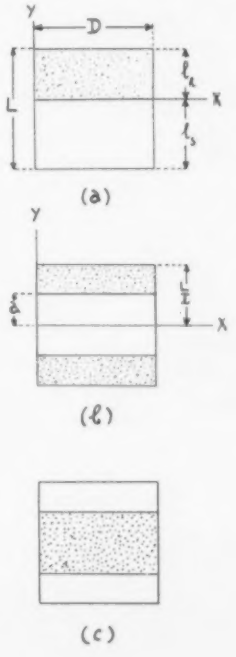


FIG. 1. Cross sections of the wave guide showing locations of hard medium.

* Graduate Fellow of the Belgian American Educational Foundation. Now with the Manufacture Belge de Lampes Electriques S. A. (Mazda), Brussels, Belgium.
† Professor of Electrical Engineering.
‡ The analysis in full, together with additional data on the structure of the modes, is included in J. Van Bladel's Ph.D. thesis, University of Wisconsin, 1950.

dependent part of a component. Thus, typically,

$$E_z=[E_z(x,y)]\cdot\epsilon^{j(\omega t+hz)}. \tag{3}$$

It also follows from Maxwell's equations that in each homogeneous medium $[E_z]$ and $[H_z]$ satisfy Helmholtz's equation

$$\nabla_{xy}^2 A+k^2A=0. \tag{4}$$

Equations (1) and (2) enable expression of the boundary conditions on $\bar{E}$ and $\bar{H}$ in terms of $[E_z]$ and $[H_z]$. The conditions at a metallic wall are $[E_z]=\partial[H_z]/\partial n=0$; the conditions at a dielectric interface, in the particular case of Fig. 1, are that

$$\frac{jh}{k_i^2}\frac{\partial}{\partial x}[E_z]-\frac{j\omega\mu_i}{k_i^2}\frac{\partial}{\partial y}[H_z], \tag{5}$$

and

$$\frac{jh}{k_i^2}\frac{\partial}{\partial x}[H_z]+\frac{j\omega\epsilon_i}{k_i^2}\frac{\partial}{\partial y}[E_z], \tag{6}$$

be continuous at the interface.

These considerations are now applied to the guide depicted in Fig. 1(a). Separation of variables in Eq. (4) and application of boundary conditions at the metallic walls yields

$$\begin{aligned}
[E_{zh}]&=A_h\sin m\pi(x/D)\sin u_h(y-l_h)/l_h,\\
[E_{zs}]&=A_s\sin m\pi(x/D)\sin u_s(y+l_s)/l_s,\\
[H_{zh}]&=B_h\cos m\pi(x/D)\cos u_h(y-l_h)/l_h,\\
[H_{zs}]&=B_s\cos m\pi(x/D)\cos u_s(y+l_s)/l_s,
\end{aligned} \tag{7}$$

where m is an integer (including zero), and u_i an auxiliary variable defined by

$$u_i=\{(\omega^2/c_i^2)-h^2-(m\pi/D)^2\}^{\frac{1}{2}},\quad i=h,s. \tag{8}$$

Applying the boundary conditions at the dielectric interface through substituting Eqs. (7) in Eqs. (5) and (6) and setting $y=0$ yields

$$\begin{aligned}
&A_h\sin u_h+A_s\sin u_s=0,\\
&B_h\cos u_h-B_s\cos u_s=0,\\
&A_h\frac{hm\pi}{D}\frac{\sin u_h}{k_h^2}+A_s\frac{hm\pi}{D}\frac{\sin u_s}{k_s^2}\\
&\qquad+B_h\frac{\omega\mu_h}{k_h^2}\frac{u_h\sin u_h}{l_h}+B_s\frac{\omega\mu_s}{k_s^2}\frac{u_s\sin u_s}{l_s}=0,\\
&A_h\frac{\omega\epsilon_h}{k_h^2}\frac{u_h\cos u_h}{l_h}-A_s\frac{\omega\epsilon_s}{k_s^2}\frac{u_s\cos u_s}{l_s}\\
&\qquad-B_h\frac{hm\pi}{D}\frac{\cos u_h}{k_h^2}+B_s\frac{hm\pi}{D}\frac{\cos u_s}{k_s^2}=0.
\end{aligned} \tag{9}$$

The determinant of the coefficients of this system of four homogeneous linear equations in A_h, A_s, B_h, B_s must be zero in order that the fields be nonvanishing. Formulating this condition yields the characteristic equation,**

$$\left(\frac{K_m u_h tgu_h}{l_h k_h^2}+\frac{u_s tgu_s}{l_s k_s^2}\right)\left(\frac{K_\epsilon u_h}{k_h^2 l_h tgu_h}+\frac{u_s}{k_s^2 l_s tgu_s}\right)+\left(\frac{hc_s m\pi}{\omega D}\right)^2\left(\frac{1}{k_s^2}-\frac{1}{k_h^2}\right)^2=0, \tag{10}$$

wherein $K_\epsilon=\epsilon_h/\epsilon_s$ and $K_m=\mu_h/\mu_s$. On determination of the cut-off frequency through imposition of $h=0$ the second term disappears and Eq. (10) can be factored into two simpler equations.

This description of the normal modes encompassed by Eqs. (7) to (10) is valid both for real and imaginary values of u_i. In the latter case trigonometric functions with imaginary arguments can be replaced by hyperbolic functions with real arguments. Proceeding thus enables distinction among trigonometric, single hyperbolic, and double hyperbolic modes whereof, respectively, neither u, only u_s, and both u_s and u_h are imaginary. We proceed now to select the lowest mode out of this ensemble.

Consideration of double hyperbolic modes can be discarded at once. For if in Eq. (10), with $h=0$, we replace u_h and u_s by $j\omega_h$ and $j\omega_s$ (where ω_h and ω_s are real), we find that the resulting equation has no roots. Thus, the double hyperbolic mode is never propagated. The contrary proves to be true of the single hyperbolic mode. Defining real auxiliary variables v_h, v_s by

$$v_h=\{(\omega^2/c_h^2)-h^2-(m\pi/D)^2l_h\}^{\frac{1}{2}}, \tag{11}$$

$$v_s=\{(m\pi/D)^2+h^2-\omega^2l_s/c_s^2\}^{\frac{1}{2}}, \tag{12}$$

the factored characteristic equation for the cut-off frequency as obtained from Eq. (10) is

$$(v_s thv_s/k_s^2l_s)-K_m v_h tgv_h/l_h k_h^2=0, \tag{13}$$

and

$$(v_s/k_s^2l_s thv_s)+K_\epsilon v_h/k_h^2l_h tgv_h=0. \tag{14}$$

Because v_s in Eq. (12) cannot be real for $h=m=0$, no single hyperbolic mode characterized by $m=0$ can be propagated. Trigonometric modes for $m=0$, i.e., independent of x and having zero E_z (H modes), can be propagated. Rewriting Eqs. (7) to (10) for $m=h=0$ gives their characteristic equation:

$$u_h tgu_h+K_\epsilon u_s tgu_s=0. \tag{15}$$

For the same integral value of $m\neq 0$, the lowest mode is single hyperbolic; for from Eq. (12) its cut-off frequency is at most $mc_s/2D$, while, from Eq. (8), it is at least this value for the trigonometric mode. Next, Eq. (13) and not (14) furnishes this lowest cut-off frequency because we note from Eq. (11) that a lower value of v_h gives a lower frequency, and Eq. (13) is satisfied by a value of v_h smaller than $\pi/2$, whereas the least value of v_h satisfying Eq. (14) is greater than $\pi/2$. Thus, the

** In Eq. (10) and following equations tg = tan and th = tanh.

lowest cut-off frequency for a given $m \neq 0$ is obtained from the intersection in a (v_h, v_s) plane of the ellipse

$$(c_h v_h/c_f)^2 + (c_s v_s/1-c_f)^2 = (\pi m L/D)^2 (c_s^2 - c_h^2), \quad (16)$$

(obtained from Eqs. (11) and (12) by setting $h=0$ and eliminating ω between the resulting two equations) with the curve

$$v_h \operatorname{tg} v_h = K_\epsilon c_f v_s \operatorname{th} v_s/(1-c_f), \quad (17)$$

(obtained from Eq. (13) by putting $h=0$). Herein $c_f = l_h/L$ measures the fraction of cross section occupied by hard medium.

The shape of the curves Eqs. (16) and (17) leads to the conclusion that smaller values of m give smaller values of v_h for the cut-off point, hence, (as we may conclude from Eq. (11)) lower cut-off frequencies. In consequence, the lowest mode is either (a) the lowest trigonometric mode for $m=0$ (an H mode), or (b) the lowest

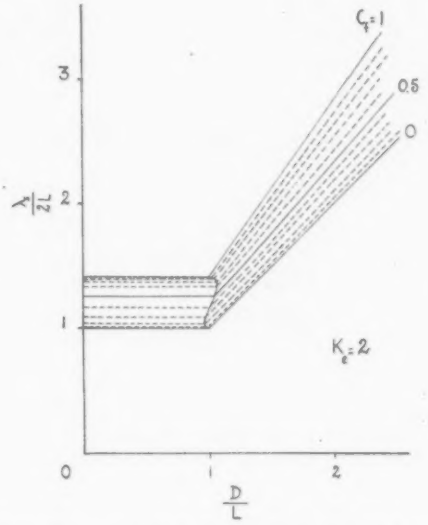


FIG. 2. Cut-off wavelength λ_s for configuration 1(a) and $K_\epsilon = 2$.

single hyperbolic mode for $m=1$, which has both E_z and H_z components (an EH mode).

The cut-off frequency of the H mode is independent of the x dimension (i.e., D). The cut-off frequency of the EH mode decreases when D increases. It is thus to be foreseen that the former mode will be the lowest for small D's, the latter for higher D's, and that there is some cross-over value of D. These conclusions are confirmed by heavy computation, the results of which are plotted in Figs. 2, 3, and 4 for media having the same magnetic permeability ($K_m = 1$) and ratios of dielectric constants K_ϵ equal to 2, 5, and 10, and for values of the parameter c_f equal to 0, 0.1, $\cdots$, 0.9, 1.

The curves of these figures express the dependence of the general cut-off frequency, measured by the dimensionless ratio $\lambda_s/2L$, in terms of the shape factors D/L and c_f. The horizontal and the inclined parts of the curves correspond, respectively, to the region where the

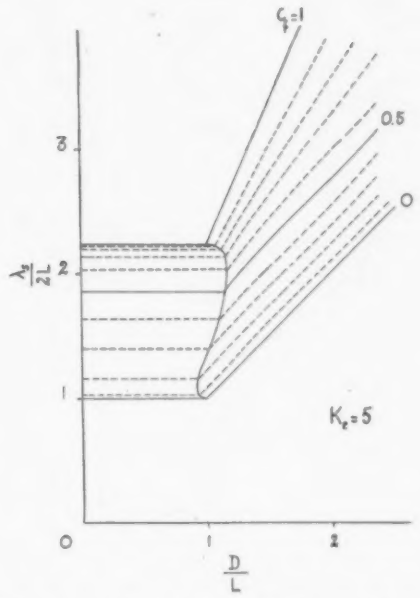


FIG. 3. Cut-off wavelength λ_s for configuration 1(a) and $K_\epsilon = 5$.

H mode and the EH mode are the lowest modes. The cross-over locus is a S-shaped curve, the tangents at the end points of which can be shown to be horizontal. The inclined parts are asymptotic to straight lines through the origin. For large values of D/L, indeed, $\lambda_s/2L$ becomes proportional to D/L, whence $\lambda_s/2D$ is independent of L. These independent values of $\lambda_s/2D$ are given in Fig. 5. The asymptotic lines which can be drawn on Figs. 2, 3, and 4 enable extension of the curves for values of D/L larger than 2, if so desired.

These figures also enable determination of whether a slab of hard medium of given volume should be put

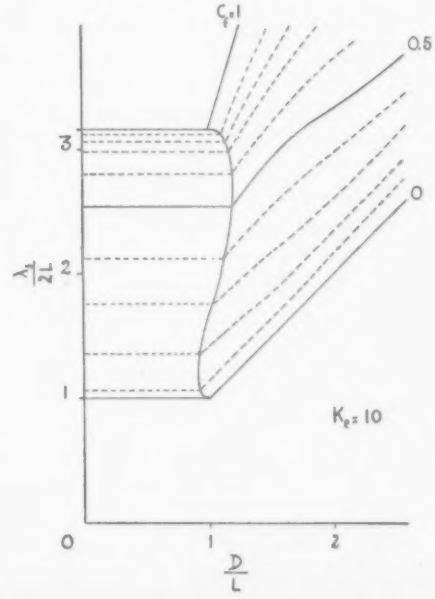


FIG. 4. Cut-off wavelength λ_s for configuration 1(a) and $K_\epsilon = 10$.

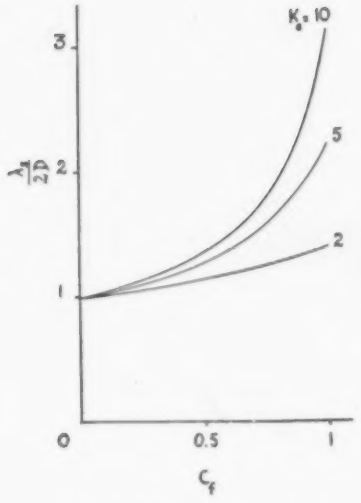


FIG. 5. Cut-off wavelength λ_c for configuration 1(a) when D/L approaches infinity.

parallel to the short or to the long side of the rectangular cross section in order to obtain maximum lowering of the cut-off frequency. A simple criterion is afforded by Fig. 6. Let $q>1$ be the ratio of the lengths of the sides of the cross section. On the curve corresponding to a given c_f we select the points P_1 and P_2 with values of D/L equal to $1/q$ and q, respectively. If point B is under A, the slab should be put parallel to the short side. Applying this criterion to Figs. 2 to 4, one finds the qualitative result that the short side should be chosen for high values of c_f and the long side for small values of c_f and that there are cross-over values of D/L and c_f.

A similar analysis can be applied to the three-layer disposition of Figs. 1(b) and (c), for which six equation similar to the four of Eq. (6) can be written. These equations involve eight unknown coefficients. Applying the boundary conditions at $y=+a$ and $y=-a$ gives eight homogeneous linear equations in these coefficients. The determinant of this system has two factors, so that when it is put equal to zero, the characteristic equation so obtained gives rise to two kinds of modes.

(a) One of the factors put equal to zero gives an equation identical to Eq. (17) except for the interchange of L with $L/2$ and l_h with a. The modes associated with this factor have antisymmetrical values for $[E_z]$ and symmetrical values for $[H_z]$ with respect to the plane $y=0$.

(b) Putting the second factor equal to zero yields modes with symmetrical values for $[E_z]$ and anti-

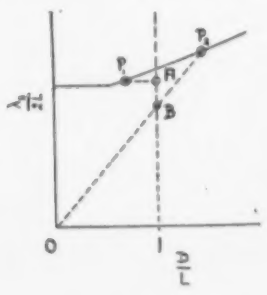


FIG. 6. Criterion for location of slab to obtain maximum lowering of cut-off frequency for configuration 1(a).

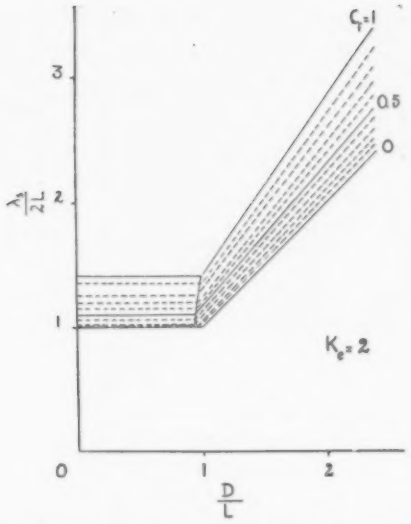


FIG. 7. Cut-off wavelength λ_c for configuration 1(b) and $K_e=2$.

symmetrical values for $[H_z]$. In particular, the lowest EH mode of this kind has the characteristic equation:

$$v_h l h v_s l g v_h / v_s = c_f K_e / (1-c_f)$$
for the arrangement of Fig. 1(b), (18)

$$v_h l h v_s l g v_h / v_s = c_f K_m / (1-c_f)$$
for the arrangement of Fig. 1(c). (19)

Comparing Eqs. (18) and (19) to Eq. (17) enables determination of whether the lowest EH mode is of type (a) or (b). As before, this mode is to be compared to the lowest H mode to determine the general cut-off frequency of the guide. It proves that the lowest EH mode is of type (a) for the arrangement of Fig. 1(b), whereas for the arrangement of Fig. 1(c) it may be either

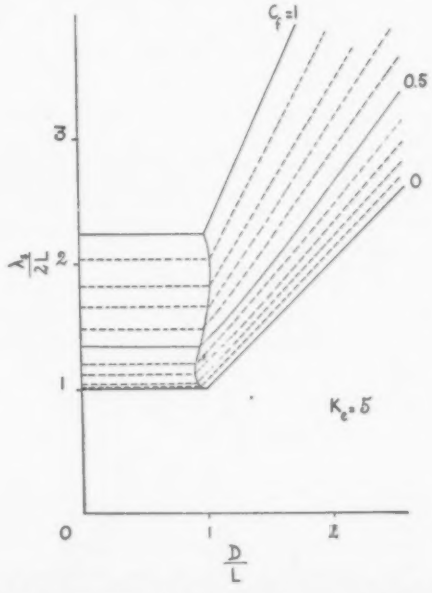


FIG. 8. Cut-off wavelength λ_c for configuration 1(b) and $K_e=5$.

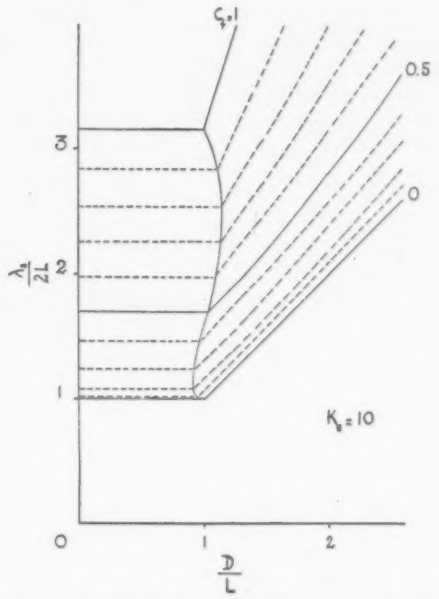


FIG. 9. Cut-off wavelength λ_1 for configuration 1(b) and $K_\epsilon=10$.

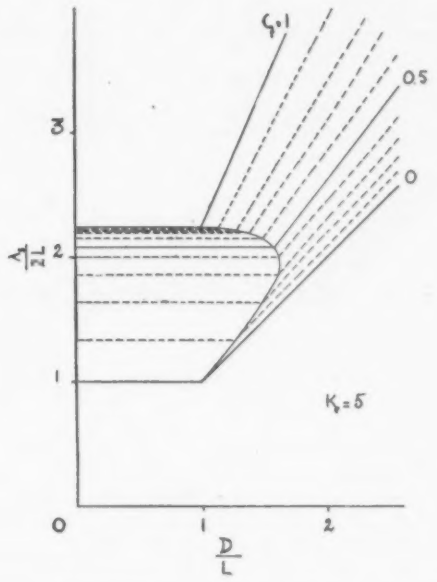


FIG. 11. Cut-off wavelength λ_1 for configuration 1(c) and $K_\epsilon=5$.

of the two, depending on the geometry of the guide. Comparison of these various modes furnishes curves for the general cut-off frequency. They are plotted in Figs. 7, 8, and 9 for the former arrangement, and in Figs. 10, 11, and 12 for the latter.

These curves are to be interpreted as above. Further, for the same values of c_f, the inclined parts of Figs. 2, 7, and 10 have the same asymptotes (and similarly for other values of K_ϵ), which the reader may draw to extrapolate the curves above $D/L=2$. From these curves it appears that the cut-off frequency is lowered a maximum amount by a given quantity of hard medium if the slab is put in the center, as in Fig. 1(c), and parallel to the short side. Various other considerations can be drawn from the graphs. In particular, the well-known results for a guide entirely filled with soft or hard medium are found for, respectively, $c_f=0$ and $c_f=1$.

APPENDIX

1. Montgomery[1] obtains, by an analysis quite different from ours, a graph of $\lambda_0/\lambda_{guide}$ of the lowest H mode for the configurations of Figs. 1(a), 1(c), and $K_\epsilon=2.45$.

2. Pincherle[2] considers the configuration of Fig. 1(a) for the case of $c_f=0.5$ and obtains equations similar to Eq. (9). However, expansion of the determinant of these equations does not give Eq. (10) for $c_f=0.5$, but rather gives Eq. (10) for $c_f=0.5$ and $m=0$.

[1] Montgomery, Dicke, and Purcell, *Principles of Microwave Circuits* (McGraw-Hill Book Company, Inc., New York, 1948), p. 385.

[2] L. Pincherle, Phys. Rev. **66**, 118 (1944).

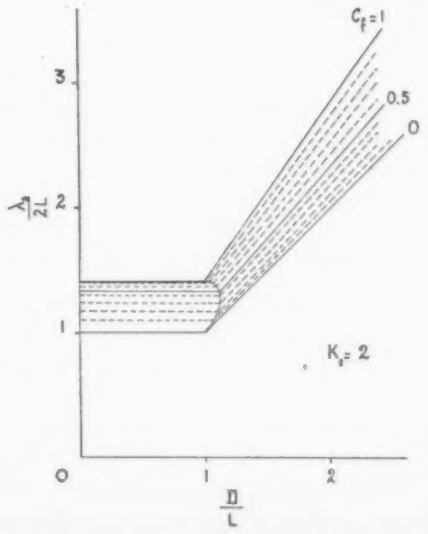


FIG. 10. Cut-off wavelength λ_1 for configuration 1(c) and $K_\epsilon=2$.

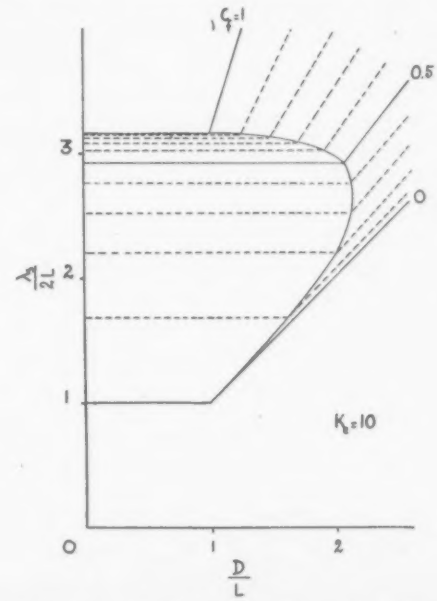


FIG. 12. Cut-off wavelength λ_1 for configuration 1(c) and $K_\epsilon=10$.

Hence, his characteristic equation is restricted to H modes, whence his later analysis based on use of this equation for values of $m \neq 0$ is invalid.

3. Moreno[3] surmises that in the special case of Fig. 1(a) for which $D/L = 2.22$ an EH mode is the lowest mode. He gives (without inclusion of details of construction) a graph of $\lambda_s/\lambda_{guide}$ for $K_t = 2.45$, $c_f = 0.5$, and $D/L = 2.22$. For $\lambda_{guide} = \infty$ (i.e., at cut-off) his chart gives $\lambda_s/2L = 2.80$. This figure is in substantial agreement with that to be derived, by interpolation, from a curve plotted from the values of 2.70, 3.20, and 3.50 given by Figs. 2–4.

[3] T. Moreno, *Microwave Transmission Design Data* (McGraw-Hill Book Company, Inc., New York, 1948), p. 195.

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

Maxwell's Demon Cannot Operate: Information and Entropy. I

L. BRILLOUIN
International Business Machines Corporation, Poughkeepsie and New York, New York

(Received September 18, 1950)

In an enclosure at constant temperature, the radiation is that of a "blackbody," and the demon cannot see the molecules. Hence, he cannot operate the trap door and is unable to violate the second principle. If we introduce a source of light, the demon can see the molecules, but the over-all balance of entropy is positive. This leads to the consideration of a cycle

Negentropy→Information→Negentropy

for Maxwell's demon as well as for the scientist in his laboratory. Boltzmann's constant k is shown to represent the smallest possible amount of negative entropy required in an observation.

I. MAXWELL'S DEMON

THE Sorting demon was born in 1871 and first appeared in Maxwell's *Theory of Heat* (p. 328), as "a being whose faculties are so sharpened that he can follow every molecule in his course, and would be able to do what is at present impossible to us···. Let us suppose that a vessel is divided into two portions A and B by a division in which there is a small hole, and that a being who *can see the individual molecules* opens and closes this hole, so as to allow only the swifter molecules to pass from A to B, and only the slower ones to pass from B to A. He will, thus, without expenditure of work raise the temperature of B and lower that of A, in contradiction to the second law of thermodynamics."[1]

The paradox was considered by generations of physicists, without much progress in the discussion, until Szilard[2] pointed out that the demon actually transforms "information" into "negative entropy"—we intend to investigate this side of the problem in a moment.

Another contribution is found in a recent paper by the present author.[3]

In order to select the fast molecules, the demon should be able to see them (see Maxwell, passage reproduced in italics); but he is in an enclosure in equilibrium at constant temperature, where the radiation must be that of the blackbody, and it is impossible to see anything in the interior of a black body. It would not help to raise the temperature. At "red" temperature, the radiation has its maximum in the red and obtains exactly the same intensity, whether there are no molecules or millions of them in the enclosure. Not only is the intensity the same but also the fluctuations. The demon would perceive radiation and its fluctuations, he would never see the molecules.

No wonder Maxwell did not think of including radiation in the system in equilibrium at temperature T. Blackbody radiation was hardly known in 1871, and it took 30 more years before the thermodynamics of radiation was clearly understood and Planck's theory developed.

The demon cannot see the molecules, hence, he cannot operate the trap door and is unable to violate the second principle.

II. INFORMATION MEANS NEGATIVE ENTROPY

Let us, however, investigate more carefully the possibilities of the demon. We may equip him with an electric torch and enable him to see the molecules. The torch is a source of radiation not in equilibrium. It pours negative entropy into the system. From this negative entropy the demon obtains "informations." With these informations he may operate the trap door and rebuild negative entropy, hence, completing a cycle:

$$\text{negentropy} \rightarrow \text{information} \rightarrow \text{negentropy}. \qquad (1)$$

We coined the abbreviation "negentropy" to characterize entropy with the opposite sign. This quantity is very useful to consider and already has been introduced by some authors, especially Schrödinger.[4] Entropy must always increase, and negentropy always decreases. Negentropy corresponds to "grade" of energy in Kelvin's discussion of "degradation of energy."

[1] The full passage is quoted by J. H. Jeans, *Dynamical Theory Gases* (Cambridge University Press, London, 1921), third edition, p. 183.

[2] L. Szilard, Z. Physik **53**, 840–856 (1929).

[3] L. Brillouin, Am. Scientist **37**, 554–568 (1949), footnote to p. 565; **38**, 594 (1950).

[4] E. Schrödinger, *What is Life?* (Cambridge University Press, London, and The Macmillan Company, New York, 1945).

We shall discuss more carefully the new cycle (1) for the demon and show later on how it extends to man and scientific observation.

The first part of the cycle, where negentropy is needed to obtain information, seems to have been generally overlooked. The second transformation of information into negentropy was very clearly discussed by L. Szilard[2] who did the pioneer work on the question.

Our new cycle (1) compares with C. E. Shannon's[5] discussion of telecommunications, which can be stated this way:

$$\text{information} \rightarrow \text{telegram} \rightarrow \text{negentropy on the cable} \rightarrow \text{telegram received} \rightarrow \text{information received.} \quad (2)$$

Shannon, however, compares information with positive entropy, a procedure which seems difficult to justify since information is lost during the process of transmission, while entropy is increased. Norbert Wiener[6] recognized this particular feature and emphasized the similarity between information and negentropy.

Our new cycle (1) adds another example to the general theory of information. We shall now discuss this problem in some detail.

III. ENTROPY BALANCE FOR MAXWELL'S DEMON

In order to discuss an entropy balance, the first question is to define an isolated system, to which the second principle can be safely applied. Our system is composed of the following elements:

1. A charged battery and an electric bulb, representing the electric torch.

2. A gas at constant temperature T_0, contained in Maxwell's enclosure, with a partition dividing the vessel into two portions and a hole in the partition.

3. The demon operating the trap door at the hole. The whole system is insulated and closed.

The battery heats the filament at a high temperature T_1.

$$T_1 \gg T_0. \quad (3)$$

This condition is required, in order to obtain visible light,

$$h\nu_1 \gg kT_0, \quad (4)$$

that can be distinguished from the background of blackbody radiation in the enclosure at temperature T_0. During the experiment, the battery yields a total energy E and no entropy. The filament radiates E and an entropy S_f,

$$S_f = E/T_1. \quad (5)$$

If the demon does not intervene, the energy E is absorbed in the gas at temperature T_0, and we observe a global increase of entropy

$$S = E/T_0 > S_f > 0 \quad (6)$$

Now let us investigate the work of the demon. He can detect a molecule when at least one quantum of energy $h\nu_1$ is scattered by the molecule and absorbed in the eye of the demon (or in a photoelectric cell, if he uses such a device).[7] This represents a final increase of entropy

$$\Delta S_d = h\nu_1/T_0 = kb \quad h\nu_1/kT_0 = b \gg 1 \quad (7)$$

according to condition (4).

Once the information is obtained, it can be used to decrease the entropy of the system. The entropy of the system is

$$S_0 = k \ln P_0 \quad (8)$$

according to Boltzmann's formula, where P_0 represents the total number of microscopic configurations (Planck's "complexions") of the system. After the information has been obtained, the system is more completely specified. P is decreased by an amount p and

$$P_1 = P_0 - p \quad \Delta S_i = S - S_0 = k\Delta(\log P) = -k(p/P_0). \quad (9)$$

It is obvious that $p \ll P_0$ in all practical cases. The total balance of entropy is

$$\Delta S_d + \Delta S_i = k(b - p/P_0) > 0, \quad (10)$$

since $b \gg 1$ and $p/P_0 \ll 1$. The final result is still an increase of entropy in the isolated system, as required by the second principle. All the demon can do is to recuperate a small part of the entropy and use the information to decrease the degradation of energy.

In the first part of the process [Eq. (7)], we have an increase of entropy ΔS_d, hence, a change ΔN_d in the negentropy:

$$\Delta N_d = -kb < 0, \quad \text{a decrease.} \quad (7a)$$

From this lost negentropy, a certain amount is changed into information, and in the last step of the process [Eq. (9)], this information is turned into negentropy again:

$$\Delta N_i = k(p/P_0) > 0, \quad \text{an increase.} \quad (9a)$$

This justifies the general scheme (1) of Sec. II.

Let us discuss more specifically the original problem of Maxwell. We may assume that, after a certain time, the demon has been able to obtain a difference of temperature ΔT:

$$\begin{array}{ll} T_B > T_A & T_B = T + \frac{1}{2}\Delta T \\ T_B - T_A = \Delta T & T_A = T - \frac{1}{2}\Delta T. \end{array} \quad (11)$$

On the next step, the demon selects a fast molecule in A [kinetic energy $\frac{3}{2}kT(1+\epsilon_1)$] and directs it into B. Then he selects a slow molecule in B, $\frac{3}{2}kT(1-\epsilon_2)$ and lets it enter into A. In order to see these two molecules,

[5] C. E. Shannon and Warren Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, Illinois, 1949).

[6] N. Wiener, *Cybernetics* (John Wiley and Sons, Inc., New York, 1948).

[7] We may replace the demon by an automatic device with a "magic eye" which opens the trap door at convenient instants of time. This is a mere problem of devising some ingenious gadget, and it does not modify the general conditions of the problem.

the demon had to use two light quanta, hence, an increase of entropy, similar to the one computed in Eq. (7):

$$\Delta S_d = 2kb, \quad b = h\nu/kT \gg 1. \tag{12}$$

The exchange of molecules results in an energy transfer

$$\Delta Q = \tfrac{3}{2}kT(\epsilon_1 + \epsilon_2) \tag{13}$$

from A to B, which corresponds to a decrease of the total entropy, on account of (11):

$$\Delta S_i = \Delta Q\left(\frac{1}{T_B} - \frac{1}{T_A}\right) = -\Delta Q\frac{\Delta T}{T^2} = -\tfrac{3}{2}k(\epsilon_1+\epsilon_2)\frac{\Delta T}{T}. \tag{14}$$

The quantities ϵ_1 and ϵ_2 will usually be small but may exceptionally reach a value of a few units. ΔT is much smaller than T, hence,

$$\Delta S_i = -\tfrac{3}{2}k\eta \quad \eta \ll 1$$

and

$$\Delta S_d + \Delta S_i = k(2b - \tfrac{3}{2}\eta) > 0. \tag{15}$$

Carnot's principle is actually satisfied.

Szilard, in a very interesting paper[2] published in 1929, discussed the second part of the process (1). He first considered simplified examples, where he could prove that additional information about the structure of the system could be used to obtain work and represented a potential decrease of entropy of the system. From these general remarks he inferred that the process of physical measurement, by which the information could be obtained, must involve an increase of entropy so that the whole process would satisfy Carnot's principle. Szilard also invented a very curious (and rather artificial) model on which he could compute the amount of entropy corresponding to the first part of the process, the physical measurement yielding the information.

In the examples selected by Szilard, the most favorable one was a case with just two complexions in the initial state,

$$P_0 = 2. \tag{16}$$

The information enabled the observer to know which one of the two possibilities was actually obtained.

$$P_1 = 1,$$

hence,

$$\Delta S_i = k(\ln P_0 - \ln P_1) = k \ln 2 < k, \tag{17}$$

since the natural logarithms of base e are used in Boltzmann's formula. Even in such an oversimplified example, we still obtain an inequality similar to (10),

$$\Delta S_d + \Delta S_i = k(b - \ln 2) > 0, \tag{18}$$

and Carnot's principle is not violated.

Szilard's discussion was completed by J. von Neumann and by P. Jordan in connection with statistical problems in quantum mechanics.[8]

[8] J. von Neumann, *Math. Grundlagen der Quantum Mechanik* (Verlag. Julius Springer, Berlin, 1932). P. Jordan, Philosophy of Science **16**, 269–278 (1949).

IV. ENTROPY AND OBSERVATION IN THE LABORATORY

The physicist in his laboratory is no better off than the demon. Every observation he makes is at the expense of the negentropy of his surroundings. He needs batteries, power supply, compressed gases, etc., all of which represent sources of negentropy. The physicist also needs light in his laboratory in order to be able to read ammeters or other instruments.

What is the smallest possible amount of negentropy required in an experiment? Let us assume that we want to read an indication on an ammeter. The needle exhibits a brownian motion of oscillation, with an average kinetic energy $(\frac{1}{2})kT$ and an average total energy

$$\bar{E}_i = kT. \tag{19}$$

G. Ising[9] considers that an additional energy

$$E_m \geqslant 4\bar{E}_i = 4kT \tag{20}$$

is required to obtain a correct reading. Let us be more optimistic and assume

$$\Delta E_m \geqslant kT. \tag{21}$$

This energy is dissipated in friction and viscous damping in the ammeter, after the physicist has performed the reading. This means an increase of entropy ΔS (or a decrease of negentropy)

$$-\Delta N = \Delta S \geqslant kT/T = k. \tag{22}$$

Boltzmann's constant k thus appears as the lower limit of negentropy required in an experiment.

Instead of reading an ammeter, the physicist may be observing a radiation of frequency ν. The smallest observable energy is $h\nu$, and it must be absorbed somewhere in the apparatus in order to be observed, hence,

$$-\Delta N = \Delta S \geqslant h\nu/T, \tag{23}$$

a quantity that must be greater than k, if it is to be distinguished from blackbody radiation at temperature T. We thus reach the conclusion that

$$-\Delta N = \Delta S \geqslant \begin{cases} k \\ h\nu/T \end{cases} \text{(whichever is greater)}, \tag{24}$$

represents the limit of possible observation.

A more compact formula can be obtained by the following assumptions: let us observe an oscillator of frequency ν. At temperature T, it obtains an average number

$$n = 1/(e^{h\nu/kT} - 1) \tag{25}$$

of quanta of energy $h\nu$. In order to make an observation, this number n must be increased by a quantity that is materially larger than n and its fluctuations; let us assume

$$\Delta n \geqslant n + 1. \tag{26}$$

[9] G. Ising, Phil. Mag. **51**, 827–834 (1926); M. Courtines, *Les Fluctuations dans les Appareils de Mesures* (Congres Intern. d'Electricité, Paris, 1932), Vol. 2.

This means at least one quantum for high frequencies ν, when n is almost zero. For very low frequencies, formula (26) requires Δn of the order of n, as in formula (21).

Observation means absorption of Δn quanta $h\nu$, hence, a change in entropy

$$-\Delta N=\Delta S=\frac{\Delta n}{T}h\nu\geqq\frac{h\nu}{T}\left[\frac{1}{e^{h\nu/kT}-1}+1\right] \tag{27}$$

or

$$-\Delta N=\Delta S\geqq k\left[\frac{x}{e^x-1}+x\right]=k\frac{x}{1-e^{-x}}x=h\nu/kT. \tag{28}$$

This formula satisfies the requirements stated in Eq. (24).

R. C. Raymond[10] recently published some very interesting remarks on similar subjects. He was especially concerned with the original problem of Shannon, Wiener, and Weaver, namely, the transmission of information through communication channels. He succeeded in applying thermodynamic definitions to the entropy of information and discussed the application of the second principle to such problems.

The limitations to the possibilities of measurements, as contained in our formulas, have nothing to do with the uncertainty relations. They are based on entropy, statistical thermodynamics, Boltzmann's constant, all quantities that play no role in the uncertainty principle.

The physicist making an observation performs the first part of process (1); he transforms negative entropy into information. We may now ask the question: can the scientist use the second part of the cycle and change information into negentropy of some sort? Without directly answering the question, we may risk a suggestion. Out of experimental facts, the scientist builds up scientific knowledge and derives scientific laws. With these laws he is able to design and build machines and equipment that nature never produced before; these machines represent most unprobable structures,[11] and low probability means negentropy? Let us end here with this question mark, without attempting to analyze any further the process of scientific thinking and its possible connection with some sort of generalized entropy.

[10] R. C. Raymond, Phys. Rev. **78**, 351 (1950); Am. Scientist **38**, 273–278 (1950).

APPENDIX

A few words may be needed to justify the fundamental assumption underlying Eq. (26). We want to compute the order of magnitude of the fluctuations in a quantized resonator. Let n_a be the actual number of quanta at a certain instant of time, and $\bar{n}$ the average number as in Eq. (26).

$$n_a=\bar{n}+m, \tag{29}$$

where m is the fluctuation. It is easy to prove the following result

$$\langle m^2\rangle_{Av}=\bar{n}(\bar{n}+1)=\bar{n}^2+\bar{n}. \tag{30}$$

For large quantum numbers, $\langle m^2\rangle_{Av}^{\frac{1}{2}}$ is of the order of magnitude of $\bar{n}$; hence, when we add n more quanta $h\nu$, we obtain an increase of energy that is only of the order of magnitude of the average fluctuations. This certainly represents the lower limit for an observation.

Formula (29) is obtained in the following way: We first compute the partition function Z (Planck's Zustandssumme) for the harmonic oscillator

$$Z=\Sigma e^{-nx}=\frac{1}{1-e^{-x}}\quad x=h\nu/kT. \tag{31}$$

Next we compute the average $\bar{n}$

$$\bar{n}=\frac{\Sigma ne^{-nx}}{z}=-\frac{1}{z}\frac{\partial z}{\partial x}=\frac{e^{-x}}{1-e^{-x}}=\frac{1}{e^x-1}, \tag{32}$$

which is our Eq. (25). Then we obtain

$$\langle n^2\rangle_{Av}=\frac{1}{z}\Sigma_1 n^2e^{-nx}=\frac{1}{z}\frac{\partial^2 z}{\partial x^2}=\frac{e^{-x}+e^{-2x}}{(1-e^{-x})^2}, \tag{33}$$

and finally

$$\langle m^2\rangle_{Av}=\langle n^2\rangle_{Av}-(\bar{n})^2=\frac{e^{-x}+e^{-2x}-e^{-2x}}{(1-e^{-x})^2}=\frac{1}{1-e^{-x}}\cdot\frac{1}{e^x-1}=\bar{n}^2+\bar{n}, \tag{34}$$

which is our formula (30). Hence, our assumption (26) is completely justified.

[11] L. Brillouin, Am. Scientist **38**, 594 (1950).

JOURNAL OF APPLIED PHYSICS VOLUME 22, NUMBER 3 MARCH, 1951

Physical Entropy and Information. II

L. BRILLOUIN
International Business Machines Corporation, Poughkeepsie and New York, New York
(Received September 18, 1950)

The laws of statistical thermodynamics are used for the definition of entropy, and it is shown that the definition of information can be reduced to a problem of Fermi-Dirac statistics or to a generalized Fermi statistics. With these definitions, the entropy of a certain message can be defined, and the information contained in the message can be directly connected with the decrease of entropy in the system.

This definition leads directly to the formulas proposed by C. E. Shannon for the measure of information, and shows that Shannon's "entropy of information" corresponds to an equal amount of negative entropy in the physical system. The physical background of the whole method is discussed and found in agreement with previous discussions.

I. INTRODUCTION

ENTROPY and probability are practically synonymous for the physicist who understands the second principle of thermodynamics as a natural tendency from unprobable to more probable structures. A new relation has been recently proposed connecting entropy with information. It is the aim of the present paper to discuss systematically the theory of information and to base it on statistical considerations that will lead to a precise definition linking "information" with a "decrease of entropy" of a system. Information is, thus, defined by the corresponding amount of negative entropy (abbreviation: negentropy). This point of view checks completely with the discussion developed in a previous paper, where it was shown that information can be obtained at the expense of negentropy and may be used to increase the negentropy of a system.

II. A SUMMARY OF THE LAWS OF STATISTICAL MECHANICS

The fundamental relation[1,2] is obtained with Boltzmann's formula[1,2] relating the entropy S of a system to its probability W:

$$S=k\ln W+S_0, \tag{1}$$

where $k=R/A=1.3\ 10^{-16}$ cgs centigrade degrees, R gas constant, A Avogadro-Loschmidt number, W probability of the system under consideration, and S_0 additional constant, ln natural logarithm base e.

$$W=P/P_T, \tag{2}$$

where P is the number of elementary structures ("complexions") corresponding to the system under consideration, P_T the total number of possible complexions for all structures of the system.

In the new quantum statistics, the notion of distinct elementary complexions obtains a rigorous definition, and Planck introduced the assumption

$$S=k\ln P, \quad S_0=k\ln P_T. \tag{3}$$

This definition satisfies the requirements of Nernst's third principle of thermodynamics and disposes of the arbitrary constant S_0. Experimental checks have been very satisfactory.[3]

Fermi-Dirac statistics applies to a distribution of N_1 similar balls into G different cells, with 0 or 1 ball per cell and never more than 1. The number of such distributions is

$$P_{FD}=G!/N_1!(G-N_1)! \qquad G\geqslant N_1. \tag{4}$$

A similar result is obtained if we discuss the distribution of N_1 black balls and N_2 white balls into G cells.

$$P_{FD}=G!/N_1!N_2! \quad \text{with} \quad N_1+N_2=G. \tag{5}$$

This is obvious, since we simply have to put the N_2 white balls into the $G-N_1$ empty cells of Fermi-Dirac.

Bose-Einstein statistics corresponds to the possible distributions of N_1 balls into G distinct cells, with any number of balls in each cell. The number of such distributions amounts to

$$P_{BE}=(N_1+G-1)!/N_1!(G-1)! \quad G\gtrless N_1. \tag{6}$$

Classical statistics represents the common limit of Fermi-Dirac or Bose-Einstein when the number of balls is very small

$$N_1\ll G.$$

FD case: $G!/(G-N_1)!$
$$=G(G-1)(G-2)\cdots(G-N_1+1)\approx G^{N_1},$$

BE case: $(N_1+G-1)!/(G-1)!$
$$=(N_1+G-1)(N_1+G-2)\cdots G\approx G^{N_1},$$

hence,

$$P_{Cl}=G^{N_1}/N_1!. \tag{7}$$

When the number p involved is large enough, the logarithm of $p!$ can be computed with Stirling's formula:

$$\ln(p!)=(p+\tfrac{1}{2})\ln p-p+\tfrac{1}{2}\ln 2\pi+\cdots\approx p(\ln p-1), \tag{8}$$

where the dots represent terms in p^{-1}, p^{-2} and so on. The simplified Stirling formula can be used for $p>100$.

[1] L. Brillouin, *Statistiques Quantiques* (Presses Universitaires, Paris, 1930); *Quantenstatistik* (Verlag. Julius Springer, Berlin, 1931), Chapters IV, V, VI.
[2] J. E. Mayer and M. G. Mayer, *Statistical Mechanics* (John Wiley and Sons, Inc., New York, 1940), pp. 92, 107.
[3] K. Darrow, Bell System Tech. J. **21**, 51–74 (1942); **22**, 108–135, 362–392 (1943).

III. ENTROPY OF A PHYSICAL SYSTEM

A physical system contains a very large number of elementary particles (molecules, atoms, electrons, photons), and we are perfectly unable to define completely the positions and velocities of all these components. All we know are the values of some "macroscopic parameters" such as pressure, volume, temperature, or total energy. This leaves an enormous amount of uncertainty about the detailed "microscopic" structure of of the system.

Let us consider a gas, for instance, and discuss the number of gas molecules to be found in a certain volume element $d\tau$, with a momentum $\mathbf{p}$ pointing into a certain volume $d\tau_p = dp_x dp_y dp_z$ of the phase extension. According to quantum mechanics, these specifications define a number of cells

$$G = (1/h^3) d\tau \cdot d\tau_p, \tag{9}$$

where h is Planck's constant.

All we want to know is the number N_1 of molecules distributed in these G cells, but we have no way to investigate the details of this distribution. Under such conditions, we compute the number of complexions P from Eqs. (4), (6), or (7), according to the type of statistics to be used, and we compute the contribution of these particles to the total entropy with Planck's formula (3), using Stirling's approximation (8).

We thus obtain the following values for the entropy:

FD
$$\frac{N}{k} = \frac{S}{k} \approx -N_1 \ln \frac{N_1}{G} - N_2 \ln \frac{N_2}{G}$$
$$N_1 + N_2 = G, \tag{10}$$

BE
$$\frac{N}{k} = \frac{S}{k} \approx -N_1 \ln \frac{N_1}{N_1+G} - G \ln \frac{G}{N_1+G}$$
$$G \gg 1, \tag{11}$$

classical
$$\frac{N}{k} = \frac{S}{k} \approx -N_1 \left(\ln \frac{N_1}{G} - 1 \right) \qquad G \gg N_1, \tag{12}$$

where N is the negentropy, a quantity whose importance was emphasized by Schrödinger and the present author. The negentropy corresponds to the "grade" of energy in Kelvin's statement about degradation of energy. In an isolated system, S must always increase; hence, N constantly decreases.

The fundamental assumption in our Eqs. (10), (11), and (12) is that we cannot distinguish between all elementary complexions and that the system under consideration is defined by the knowledge of the number of balls N_1 and the number of cells G. The classical formula (12) leads directly to Boltzmann's theory of an ideal gas:

$$S = kG(-p_1 \ln p_1) + kN_1 \qquad p_1 = N_1/G, \tag{13}$$

where p_1 is the density of particles per cell, G the number of cells.

The total entropy of the system is obtained by taking the sum for all possible volume elements and all momentum values,

$$S_{\text{total}} = k \sum G(-p_1 \ln p_1) + kN. \tag{14}$$

In the last term, N represents the total number of particles in the gas and is a constant of the system; hence, the result

$$S_{\text{total}} = \sum G \cdot H_1 + S_0$$
$$H_1 = -k p_1 \ln p_1, \tag{15}$$

where H_1 is the quantity playing the fundamental role in Boltzmann's famous "H" theorem. The most probable distribution is the one that makes S_{total} a maximum.

IV. GENERALIZED FERMI-DIRAC STATISTICS

We shall have to consider the following problem: given N_1 balls of type 1, N_2 balls of type 2, $\cdots N_n$ balls of type n, and a number of cells,

$$G = N_1 + N_2 \cdots + N_n = \sum_{i=1}^{n} N_i, \tag{16}$$

what is the number P_{GF} of distinct possible distributions of these balls into the G cells with no more than one ball per cell?

$$P_{GF} = \frac{G!}{(N_1! N_2! \cdots N_n!)} = \frac{G!}{\prod_i (N_i!)}. \tag{17}$$

This represents a generalization of our Fermi-Dirac formula (5).

Another problem is important to discuss. What is the total number P_t of possible distributions when the numbers $N_1, N_2, \cdots N_n$ are not specified but can take any set of values corresponding to a given total G in Eq. (16)?

Let us first consider the case of two balls and Fermi-Dirac statistics. We have to sum all possible N_1 values in Eq. (4)

$$P_t = \sum_{N_1=0}^{N_1=G} \frac{G!}{N_1!(G-N_1)!}. \tag{18}$$

This can be compared with the binomial formula

$$(\alpha+\beta)^G = \sum_{N_1=0}^{G} \frac{G!}{N_1!(G-N_1)!} \alpha^{N_1} \beta^{G-N_1}. \tag{19}$$

Taking

$$\alpha = \beta = 1,$$

we obtain

$$P_t = 2^G. \tag{20}$$

We may have to correct this result if we do not want to distinguish between a distribution of N_1 black and N_2 white balls, and another distribution with N_2 black and N_1 white balls. The result should then be halved

$$P_{tc} = \tfrac{1}{2} 2^G. \tag{21}$$

Formula (20) is obvious. For each cell we have 2 choices (black or white), hence, 2^G for G cells. Formulas (18) and (19) show the connection with our previous computation.

Let us repeat this for n different types of balls:

$$P_t=\sum_{N_1N_2\cdots N_n}\frac{G!}{\prod_i N_i!}, \tag{22}$$

and the polynomial formula reads

$$(\alpha+\beta+\cdots+\nu)^G=\sum\frac{G!}{\prod_i N_i!}\alpha^{N_1}\beta^{N_2}\cdots\nu^{N_n}. \tag{23}$$

Taking $\alpha=\beta=\cdots=\nu=1$, we obtain

$$P_t=n^G \tag{24}$$

or

$$P_{to}=(1/n!)n^G. \tag{25}$$

If we correct P_t for not distinguishing between cases where the n types of balls have been simply interchanged, Eqs. (24) and (25) are also pretty obvious.

V. PROBABILITIES APPLIED TO INFORMATION: A SIMPLE EXAMPLE

Let us consider a simplified model of a telegram, where only dots and blanks would be used. We have G positions available, with a total of N_1 dots and N_2 blanks;

$$N_1+N_2=G. \tag{26}$$

This general statement may apply to electric signals on a cable, pulses on a mercury delay line, magnetization on magnetic tape, and many similar problems.

Each dot should, in principle, correspond to a pulse having the length of one of our G "positions" and a given intensity; but we may allow for some small admissible variations in the shape, intensity, and length of these pulses. This leads to the introduction of the number P_1 of different possible types of pulses that may be permitted. These P_1 types will play the role. of P_1 elementary "complexions" discussed in Sec. II In a similar way, we introduce P_2 possible types of blanks. The situation would be the same if we were to use, instead of dots and blanks, two different types of pulses (positive and negative for instance).

We now want to consider different statistical problems.

A. We may take the point of view of Sec. III and assume that we are not interested in any specific distribution of dots and blanks but only in the numbers N_1, N_2, G, and we may ask about the total number of different distributions that can be obtained, once N_1, N_2, and G are given. The answer is obtained from Fermi-Dirac statistics. We may have either a dot or a blank on each of the G positions.

Using formula (5) and allowing for the P_1 and P_2 different types of dots and blanks, we find

$$P_{N1N2}=P_1^{N_1}P_2^{N_2}(G!/N_1!N_2!) \tag{27}$$

for the total number of different distributions.

B. If we are interested in the distribution of dots and blanks, because our telegram corresponds to one specific distribution, the number of permissible combinations is only

$$P'_{N1N2}=P_1^{N_1}P_2^{N_2}. \tag{28}$$

C. Comparing Eqs. (27) and (28), we notice that once the numbers N_1 and N_2 are given, the probability W_{N1N2} of a certain message can be defined as the ratio of the number P'_{N1N2} of cases to the total possible number P_{N1N2}

$$W_{N1N2}=P'_{N1N2}/P_{N1N2}=N_1!N_2!/G! \tag{29}$$

or

$$P'_{N1N2}=W_{N1N2}\cdot P_{N1N2}. \tag{30}$$

Let us now try to define entropy for such a problem.

Physical entropy, as defined in Sec. III, corresponds to problem A and Eq. (27). Let us assume our telegram to be long enough. Using the Stirling formula, we obtain an entropy similar to that obtained in (10)

$$\begin{aligned}S_{\text{phys}}&=k\ln P_{N1N2}\\&=k[-N_1\ln(N_1/P_1G)-N_2\ln(N_2/P_2G)]\\&=kG[-p_1\ln(p_1/P_1)-p_2\ln(p_2/P_2)]\end{aligned}$$
$$p_1=N_1/G,\quad p_2=N_2/G. \tag{31}$$

Here p_1 and p_2 represent the average density of dots or blanks per cell, and the density of physical entropy per cell is

$$s_{\text{phys}}=S_{\text{phys}}/G=k[-p_1\ln(p_1/P_1)-p_2\ln(p_2/P_2)]. \tag{32}$$

The entropy of a given message is obtained from Eq. (28).

$$S_m=k\ln P'_{N1N2}=k(N_1\ln P_1+N_2\ln P_2), \tag{33}$$

and its density per cell amounts to

$$s_m=S_m/G=k(p_1\ln P_1+p_2\ln P_2). \tag{34}$$

The *information* contained in the message can be defined as

$$I=S_{\text{phys}}-S_m=-k\ln W_{N1N2}>0, \tag{35}$$

and its density per cell is

$$i=I/G=k[-p_1\ln p_1-p_2\ln p_2]. \tag{36}$$

According to these definitions, information plays the role of negentropy, since

$$S_m=S_{\text{phys}}-I. \tag{37}$$

The entropy of the physical system carrying the message is smaller than the maximum physical entropy, and the difference represents the information I contained in the telegram.

VI. PROBABILITY AND ENTROPY FOR A MESSAGE USING n DIFFERENT TYPES OF PULSES

We may proceed along similar lines for a problem where n different types of pulses might be used instead of just two different pulses as in the preceding example. Let $N_1 N_2 \cdots N_n$ be the number of pulses of different types, and G the total number;

$$N_1+N_2\cdots+N_n=G. \tag{38}$$

Let P_1, P_2, $\cdots P_n$ be the number of complexions for each type of pulse. We again consider different statistical problems.

A. We may assume that we are not interested in any specific distribution of the pulses, but we care only for the numbers of pulses of each type. Using Eq. (17) and taking into account the $P_1\cdots P_n$ possibilities, we obtain a total number of different distributions:

$$P_{N1\cdots Nn}=P_1^{N_1}P_2^{N_2}\cdots P_n^{N_n}(G!/N_1!N_2!\cdots N_n!). \tag{39}$$

This formula generalizes our previous Eq. (27).

B. A specific telegram corresponds to a certain distribution of the pulses and can be realized in

$$P'_{N1\cdots Nn}=P_1^{N_1}P_2^{N_2}\cdots P_n^{N_n} \tag{40}$$

different ways.

C. Comparing (39) and (40), we define the probability of our telegram

$$W_{N1\cdots Nn}=(P'_{N1\cdots Nn})/(P_{N1\cdots Nn}) =(N_1!N_2!\cdots N_n!)/G! \tag{41}$$

The *physical entropy* is again given by

$$S_{\text{phys}}=k\ln P_{N1\cdots Nn} =k\left[-N_1\ln\frac{N_1}{P_1G}-N_2\ln\frac{N_2}{P_2G}\cdots-N_n\ln\frac{N_n}{P_nG}\right], \tag{42}$$

and its density per cell amounts to

$$s_{\text{phys}}=\frac{S_{\text{phys}}}{G} =k\left[-p_1\ln\frac{p_1}{P_1}-p_2\ln\frac{p_2}{P_2}\cdots-p_n\ln\frac{p_n}{P_n}\right], \tag{43}$$

where $p_i=N_i/G$ represents the average density of pulses of type i per cell.

The entropy of a given message is obtained from Eq. (40):

$$S_m=k\ln P'_{N1\cdots Nn}=k[N_1\ln P_1+\cdots N_n\ln P_n], \tag{44}$$

and its density per cell is

$$s_m=\frac{S_m}{G}=k[p_1\ln P_1\cdots+p_n\ln P_n]. \tag{45}$$

The *information* contained in the message is the entropy difference

$$I=S_{\text{phys}}-S_m=-k\ln W_{N1\cdots Nn}>0, \tag{46}$$

with a density per cell

$$i=\frac{I}{G}=k[-p_1\ln p_1\cdots-p_n\ln p_n], \tag{47}$$

and information means negentropy, as in the previous example, Eq. (37).

VII. A COMPARISON WITH SHANNON'S DEFINITIONS

In his fundamental discussion of the theory of information,[4] Shannon links the definition of information with the concept of entropy in the following way (reference 4, p. 18): "Can we define a quantity which will measure how much information is produced in a process, or better at what rate information is produced?

"Suppose we have a set of possible events, whose probability of occurrence are p_1, p_2, p_n. These probabilities are known, but that is all we know concerning which event will occur. Can we find a measure of how much *choice* is involved in the selection of event, or of how uncertain we are of the outcome?"

Shannon is considering the time as an independent variable. The "rate at which information is produced" corresponds to the density of information per unit of time, We used the density per unit of length of a signal. Both quantities are proportional to each other:

$$i_t=i\cdot g, \tag{48}$$

where i_t represents density of information per unit time, i the density of information per "cell" (length of a pulse), g the number of "cells" per unit time.

Shannon discusses some general properties required from a definition of this information density and finally assumes (reference 4, p. 19) that it must be represented by

$$H=-K\sum_{i=1}^{n}p_i\ln p_i \tag{49}$$

with an arbitrary constant K.

This is exactly the definition we obtained from statistical considerations at the end of the preceding section [Eqs. (47) and (48)].

$$i_t=i\cdot g=-k\cdot g\sum p_i\ln p_i. \tag{50}$$

Hence, our statistical discussion justifies Shannon's definition and shows that his "entropy of information" corresponds to a negative term in the physical entropy of the system carrying the message.

The question of sign required a detailed discussion of the statistics involved, on account of the misleading

[4] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, Illinois, 1949).

similarity between Boltzmann's formula [Eqs. (15) or (31)] and our formula (34), which involve opposite signs.

Shannon discusses (reference 4, p. 23) the probability problem for a source of information and computes the average density of information for a long message containing N intervals. This number N, introduced by Shannon, corresponds to our "number of cells" G in Secs. III, IV, V, and VI.

Shannon uses classical statistics instead of the Fermi-Dirac scheme. The discussions which gave rise to both the Fermi-Dirac and the Bose-Einstein statistics have proved, many years ago, that the so-called classical statistics was correct only for infinitely small cells (yielding very large numbers for G) and could not be applied to cells of finite size. We showed in Sec. III, Eq. (7), how classical statistics is obtained in the limit $G \gg N_1$.

At another place (reference 4, p. 27, theorem 7) Shannon proves that a transducer must *decrease* (or at best maintain constant) the *entropy of information*. A transducer plays in this theory a similar role to a transformation in physical thermodynamics. An irreversible transformation always *increases physical entropy*, and a reversible transformation maintains a constant entropy. This comparison shows again that Shannon's "entropy of information" corresponds to negative physical entropy. It also proves that the "nonsingular" transducer, that keeps entropy constant, is a "reversible" transducer.

N. Wiener[5] recognized that information means negative entropy. When a telegram is transmitted, some errors are introduced, part of the information is lost, and by the same token, the entropy of the system is increased. Hence, information compares with negentropy.

Wiener uses this remark to reverse the sign of Shannon's formula. We cannot agree on this point. Shannon's formula is all right, only the interpretation of the formula is more involved.

VIII. GENERAL DEFINITIONS

A summary of our viewpoint may be useful at this place and will explain the position of the whole question. Let us consider a physical system, a gas for instance. Its state can be defined by giving the number N of gas molecules, the volume V, and the total energy E. With these data we may compute the total number P of possible elementary "complexions" that give us the value of the entropy S of the system, using the Boltzmann-Planck Eq. (3). Once we know the function $S(N, V, E)$, we obtain the temperature

$$T = (\partial E/\partial S)_{V \text{ constant}} \tag{51}$$

and the pressure

$$P = -(\partial E/\partial V)_{S \text{ constant}}. \tag{52}$$

[5] N. Wiener, *Cybernetics* (John Wiley and Sons, Inc., New York, 1948), p. 76, equation (3.05).

The very scanty information we have about the system when N, V, E are given makes it possible to find a very large variety of permissible structures for the whole system. It leaves us with a great freedom of choice, which means a large P and a large entropy S.

If, by any chance, we are able to obtain some additional information about the structure of the system, our freedom of choice is decreased and so is the entropy. Assume, for instance, that at a certain instant of time all the gas molecules happen to be located in a smaller volume $V_1 < V$, then the entropy at this time is $S_1 < S$, but if our molecules are free to move through a vessel of volume V, the preceding situation will last only a very short time and the molecules will very soon be scattered over the whole volume V, yielding the higher entropy S.

This example is just a special instance of a general rule: any additional information about the system under consideration corresponds to a decrease in the total number of complexions P. We have to reject all the complexions that would not satisfy the additional conditions contained in our information. Hence, the total entropy decreases whenever we happen to have some special information about the structure of our physical system. We choose this entropy decrease as the physical measure of information.

$$I = -\Delta S = \Delta N, \quad S, \text{ entropy}; \quad N, \text{ negentropy}. \tag{53}$$

This situation will always be unstable. After a while, the information loses its value, the system does not satisfy any more the corresponding conditions, the entropy increases to its maximum value. For some type of processes this natural evolution may be very fast; for other types, it may be relatively slow. The second principle of thermodynamics requires that S should increase to its maximum value, but it does not say anything about the speed of this recovery.

Systems of importance, for telecommunications (or for memory devices in computing equipment), are those which possess a certain number of relatively stable structures, namely, structures that will persist for a long time before they eventually get destroyed and the entropy goes back again to its maximum. For such structures, the information I is maintained for an appreciable length of time, and the system can be used for memory or communication.

Let us return to our example with the gas in a container V. Assume plane parallel walls and an initial situation where a plane wave is moving from one wall to the opposite one; the wave will go on propagating from left to right, be reflected, then come back from right to left, be reflected on the first wall, and so on for a certain time, until viscosity effects finally destroy the wave and change its energy into heat motion. The information contained in the wave persists for a practical length of time before it eventually disappears. The system can be used for storage of information (replace the gas by mercury and you have the mercury

delay line memory system) or for communication. The case of electromagnetic waves propagating along a cable is very similar. The wave may be picked up by a receiver at the end of the cable, or it may be reflected and propagate back and forth until it finally dies out by ohmic resistance. At any rate, when information disappears, the whole system goes back to its maximum entropy value.

Transmission or storage of information is associated with the temporary existence of the system in a state of lower entropy. The decrease of entropy can be taken as a measure of the amount of information [Eq. (53)].

The point of view developed in this section checks completely with the discussion given in a previous paper[6] about information and entropy in physical problems.

IX. ENTROPY OF A SYSTEM WITH CONSTRAINTS

If we impose some constraints upon a physical system, its entropy is decreased. A constraint represents a special kind of limitation or of information about the structure of the system.

In our discussion of Secs. V and VI, we used a typical constraint when we stated that we were only considering certain types of pulses of given length and shape. What we called the *physical entropy*, under such conditions, was a quantity defined similarly to the actual entropy but maintained at a lower value by the constraints.

This was done in order to distinguish between two different parts in the total information:

A. the part corresponding to the technical organization of the system of transmission—type of pulses, their duration, their number, and

B. the information contained in the telegram itself, when a specific succession of pulses has been selected.

Part A is of interest for the technician in charge of the transmission system. Part B is the only one of importance to the customer and represents the actual information value.

We discussed problem B extensively. Let us now say a few words about problem A. We consider, for simplicity's sake, the situation discussed in Sec. V with dots and blanks. We obtained in Eq. (31) a physical entropy,

$$S_{\text{phys}}=kG[-p_1\ln(p_1/P_1)-p_2\ln(p_2/P_2)], \quad (31)$$

that corresponds to a constraint, where dots and blanks are supposed to have all the same length, and the total number of cells (or intervals) is G. The length of the cells required for message transmission will be very much larger than the actual physical cells, where the limitation is to be found only in Planck's constant h. Hence, a real physical problem will first introduce a number of cells,

$$G_{rp}=qG \quad q\gg 1. \quad (54)$$

In addition to this, the permissible types of dots and blanks P_1 and P_2 will also be enormously increased, while the densities p_1 and p_2 will remain approximately constant, since this means a conservation of the total energy of the pulse system as a whole. We assume in the real physical problem

$$P_{1r}=rP_1 \quad P_{2r}=rP_2 \quad r\gg 1; \quad (55)$$

and we obtain, for the real physical problem,

$$\begin{aligned} S_{r\ \text{phys}}&=kG_{rp}[-p_1\ln(p_1/P_{1r})-p_2\ln(p_2/P_{2r})]\\ &=kqG[-p_1\ln(p_1/P_1)-p_2\ln(p_2/P_2)\\ &\qquad +(p_1+p_2)\ln r];\\ S_{r\ \text{phys}}&=q[S_{\text{phys}}+kG\ln r], \quad q\gg 1 \quad r\gg 1, \end{aligned} \quad (56)$$

since $p_1+p_2=1$.

The real physical entropy of the system is very much larger than the physical entropy with constraints. The negentropy,

$$N_A=(S_{r\ \text{phys}}-S_{\text{phys}}), \quad (57)$$

represents the price we have to pay in order to obtain a readable message, which our receiving station can interpret correctly. This part of the negentropy is of interest only to the engineer who designs the communication system. As a matter of fact, it will not be of such importance as the relative stability of the signals employed for the transmission. When we send a message along the transmission channel, its initial physical entropy is S_{phys}, and it progressively disintegrates on account of viscosity or ohmic resistance or of other causes for dissipation. Finally, the physical entropy will reach its maximum value, Eq. (56). The most important problem for the engineer is not the actual difference of entropy, Eq. (57), but the time constant involved.

As a conclusion, we may say that a certain entropy change is essential in the definition of information (problem B). In the discussion of the quality of a transmission or memory device (problem A), the relative stability (life time) is the important feature, while the change in entropy represents a secondary character.

In all these discussions, what we called the real physical entropy $S_{r\ \text{phys}}$ represents the *final* entropy increase of the system after all the energy contained in the signals has been dissipated into heat. The initial entropy of the system, before the signals have been emitted, does not appear anywhere in our theory. It would simply correspond to an additional constant S_0 in all the entropy formulas.

[6] L. Brillouin, J. Appl. Phys. **24**, 334 (1950).

A Metallographic Study of Diffusion Interfaces*

G. C. KUCZYNSKI† AND B. H. ALEXANDER
Sylvania Electric Products, Inc., Bayside, New York
(Received September 5, 1950)

By sintering wires to flat metallic blocks, diffusion interfaces were obtained between the following pairs of metals: Cu-Ni, Au-Ag, Cu-Ag, Cu-Fe, Fe-Ni, Cu-Zn, Ni-Zn, and 70/30 brass-Cu. The interfaces were subsequently investigated metallographically. Grooves were observed in the lower melting point component in the vicinity of the sintered interfaces. This is attributed to the faster diffusion rate in the lower melting point metal.

DIFFUSION studies between two metals are most commonly made by using plane interfaces. The geometry of such interfaces has not been observed to change during diffusion, except possibly at the extreme edges.[1] In the present work diffusion couples were made where the interface was not planar, and marked changes in geometry occurred during the course of diffusion. These observations should be of interest not only because they may give some indication as to diffusion mechanism but also because of their possible practical implications.

When a small wire of a metal is sintered to a surface of the same metal, the bond formed is uniform and symmetrical as can be seen in Fig. 1. In practically all cases of this kind a grain boundary is formed across the shortest neck width, as would be expected. The formation of such bonds has been explained on the basis of volume diffusion.[2,3] The process can be described as a spheroidization of the sharp filets on either side of the contact between the wire and the surface.

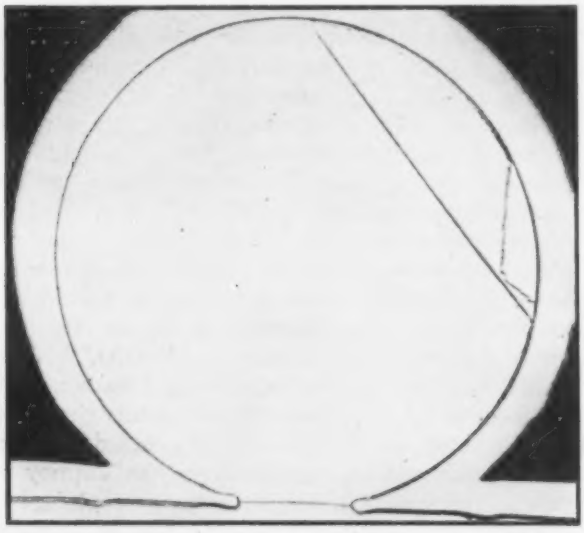


FIG. 1. Cross section of 10-mil diameter copper wire which has been sintered to a copper surface 25 hr at 1050°C. The specimen was electroplated with nickel before polishing, ×187.5.

The symmetrical bond formation shown in Fig. 1 is seldom maintained when the wire is a different metal than the surface. In most cases of the latter kind, the shortest width of the bond is not maintained at the point of original contact of the wire and the surface.

The following systems were investigated by winding small wires (5–10-mil diameter) of one metal around ¼-inch cylinders of the other metal and heating such specimens for various periods of time: Cu-Ni, Au-Ag, Cu-Ag, Cu-Fe, Fe-Ni, Cu-Zn, Ni-Zn, and 70/30 brass-Cu. (A hydrogen or argon atmosphere was used during the diffusion anneals except for the specimens containing brass or zinc, which were heated in vacuum.) After heating, the specimens were electroplated with nickel if necessary to prevent damage during metallographic preparation, and were then mounted in Bakelite and cut parallel to the axis of the ¼-inch cylinder in order to show the cross section of the wire which had bonded to the surface of the cylinder.

Cu-Ni

The interfaces formed between Cu-Ni and Ni-Cu couples are shown in Figs. 2 and 3. It is to be noted that a groove has formed in both cases on the copper

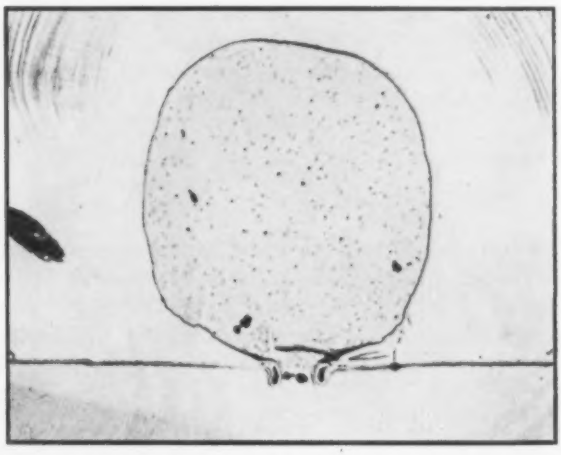


FIG. 2. Cross section of 5-mil nickel wire on copper, 1 hr at 995°C. The sample was nickel plated after heating, but the nickel did not completely fill the grooves which had formed in the copper, ×300.

* This work has been supported in part by the AEC, Contract No. AT-30-1 GEN 367.
† Now at NACA, Lewis Flight Propulsion Laboratory, Cleveland, Ohio.
[1] W. A. Johnson, AIME **147**, 331 (1942).
[2] G. C. Kuczynski, J. Metals **1**, 169 (1949).
[3] G. C. Kuczynski, J. Appl. Phys. **21**, 632 (1950).

FIG. 3. Cross section of 5-mil copper wire on nickel, 64 hr at 900°C. After this prolonged heating, the "digging" has progressed more than halfway through the copper wire, which in turn has become somewhat distorted, ×450.

side of the bond. This is characteristic of this type of diffusion couple, but the junction is very irregular in many cases and appears to be influenced somewhat by the atmosphere. When the specimens were heated in hydrogen, the "digging" into the copper was more pronounced than when they were heated in argon.

Au-Ag

When gold wire was wound on a silver cylinder and heated, the "digging" occurred into the silver (Fig. 4); but the effect was less pronounced than in the copper-nickel specimens.

Fe-Ni

In the systems already mentioned there was complete solid solubility between the two metals forming the diffusion couple, and concentration changes near the bond could not readily be inferred. In the Fe-Ni system, there is complete solid solubility at high temperature; but the iron rich portion will transform on cooling to

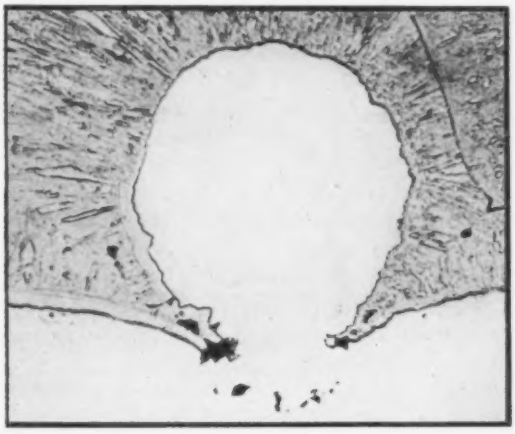


FIG. 4. Five-mil gold wire on silver, 6 hr at 900°C. Here the "digging" occurred in the silver. The holes below the junction suggest that a smaller groove had formed earlier and has subsequently been covered over, ×225.

FIG. 5. Five-mil nickel wire on iron, 6 hr at 1185°C. The "digging" in this case is in the nickel. These grooves have not been filled in by the nickel plate subsequent to heating. Note that the nickel has spread out considerably over the iron surface, ×300.

room temperature, and the phase boundary thus formed can easily be observed. In the wire-cylinder diffusion specimens in this system, the "digging" occurred in the nickel-rich phase, Figs. 5 and 6. These photo-micrographs also demonstrate that the nickel is diffusing along the surface of the iron more rapidly than into the volume. In some, but not all, cases the nickel diffused preferentially along a grain boundary in the iron.

Cu-Ag

Copper wire wound on silver gave a very irregular bond in most cases, but there was a slight tendency for "digging" into the silver. These specimens were similar in appearance to the Cu-Fe couples described below.

Cu-Fe

The copper-iron system consists of two phases at higher temperatures; and in this system, also, the progress of diffusion can be observed qualitatively by the movement of the boundary between the iron-rich and the copper-rich phases. When iron wire was wound on copper cylinders and heated, "digging" was observed

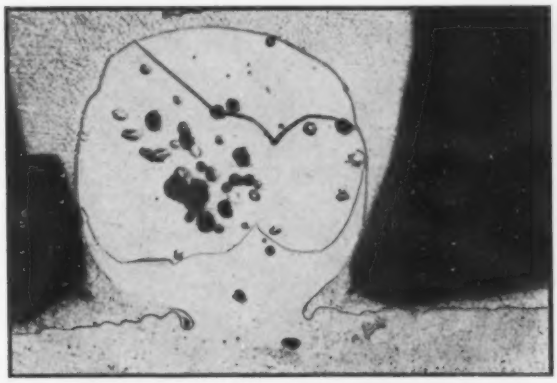


FIG. 6. Five-mil iron wire on nickel, 6 hr at 1185°C. Beside diffusing over the iron surface, the nickel has apparently penetrated more rapidly into a grain boundary in the iron, ×300.

FIG. 7. Five-mil iron wire on copper, 65 hr at 1013°C. The grooves have formed in the copper, but there is much less evidence of surface diffusion than in the iron-nickel couples, ×300.

in the copper-rich phase, Fig. 7. There was a slight tendency for the copper to diffuse faster along the surface of the iron, but this was not nearly as marked as in the Fe-Ni system. There is indirect evidence that the copper has penetrated the iron wire to a larger extent, presumably by a combination of surface and volume diffusion, in that the grain size of the iron wire after diffusion is quite small. The same iron wire when heated in the same manner without the presence of copper showed a very large grain size. Thus, it appears that the presence of copper is responsible for the small grain size in the present case. There was no observable tendency for the copper to penetrate the iron grain boundaries preferentially, but the boundaries visible in Fig. 7 may have formed on cooling through the $\gamma \rightarrow \alpha$-transformation.

The reverse case of this system (copper wire wound on iron cylinders) showed another unusual feature. In the normal case there was "digging" into the copper wire, and also a tendency for the copper to spread out slightly over the surface of the iron, Fig. 8. In many cases, however, it appeared that the junction as a whole was "sinking" down into the iron surface, Fig. 9. In a few cases of this kind it appeared as if the "sinking" were associated with the presence of an iron grain boundary at the edge of the neck as in Fig. 10, where only one side of the junction is displaced.

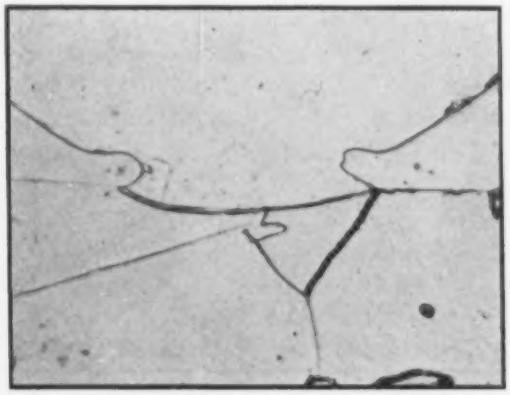


FIG. 8. Cross section of the junction between 5-mil copper wire on iron, 65 hr at 1013°C.

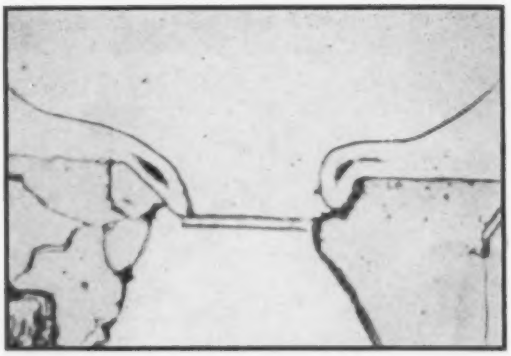


FIG. 9. This couple was prepared in the same manner as the one shown in Fig. 8, although in this case it appears that the junction has sunk into the iron base.

Cu-Zn

The progress of the bond formation between copper wire wound on a zinc cylinder is shown in Figs. 11 through 14 for two different temperatures. It is to be noted in this case that a groove forms on either side of the bond, digging down into the zinc surface, and that this groove always terminates at the junction between the ϵ and η-phases. At longer times of heating this groove widens and becomes deeper.

Attempts to bond zinc wire to copper surfaces were not informative as to the geometry of the bond, since it was always broken during preparation or during cooling. Longer times had to be used to obtain a strong enough bond; and this resulted in a general penetration over the entire surface of the copper, Fig. 15.

Ni-Zn

Annealed specimens of nickel wire wound on zinc cylinders resulted in a structure which was initially similar to the copper-zinc system. There is a decided

FIG. 10. Same as Fig. 9, showing a region where only one side of the junction is displaced.

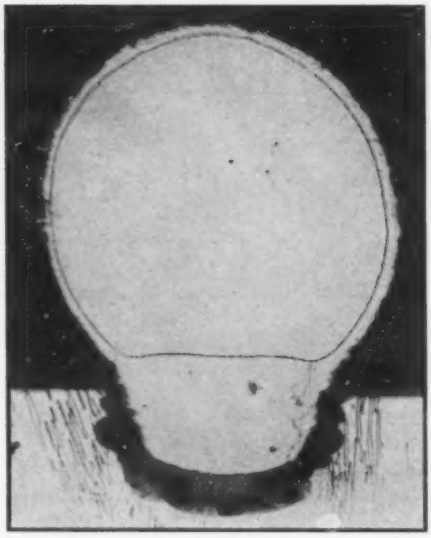


FIG. 11. Ten-mil copper wire on zinc, 1 hr at 368°C. The core of the copper wire is now α-brass and is surrounded by a very thin layer of β-brass and then a thick layer of γ-brass. The dark etching phase near the bottom is ϵ and the groove terminates at the junction between the ϵ-phase and the Zn rich η-phase forming the base, ×187.5.

tendency toward columnar growth of a Ni-Zn phase from the surface of the nickel wire, Fig. 16. At longer periods of time this growth has become very irregular, Fig. 17, and finally attains an orchid-like appearance, Fig. 18.

70/30 BRASS-Cu SYSTEM

Figures 19 and 20 represent the cross sections of 0.010-inch diameter copper wire sintered to brass cylinders. The bond was poorly formed in many cases; it

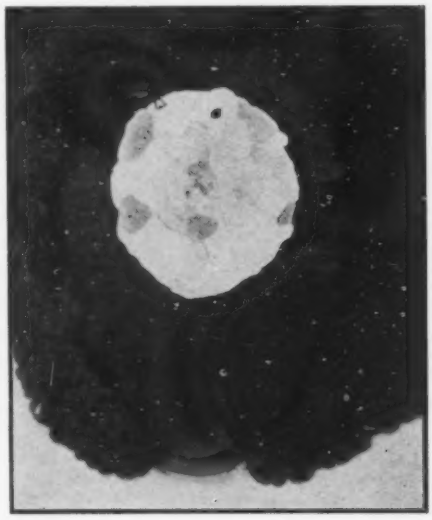


FIG. 12. Ten-mil copper wire on zinc, 24 hr at 368°C. The core of the copper wire has now been converted to γ-brass, surrounded by the dark ϵ-phase. The grooves have now become very wide, ×75.

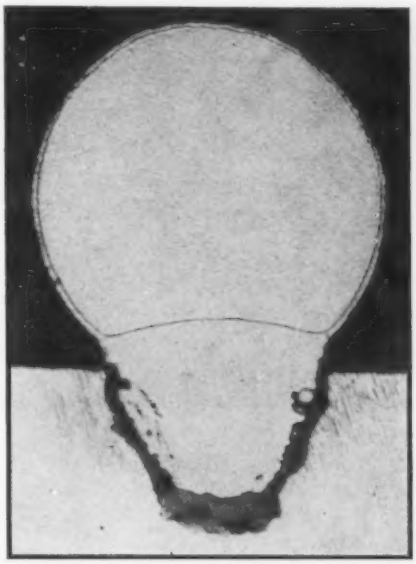


FIG. 13. Ten-mil copper wire on zinc, 8 hr at 300°C. The phases present here are the same as those shown in Fig. 11, ×187.5.

appeared that oxide inclusions were interfering with the normal growth of the bond. It is to be observed that zinc evaporates from the brass and deposits on the surface of the wire, and subsequently diffuses inward. The progress of diffusion can be watched on cross section because almost pure copper metal in the center of the wire etches darker than the zinc rich outer layer. It should be noted that there is no apparent alteration of the almost circular etching boundary of the copper wire on the side facing the brass, indicating that the transport is chiefly by evaporation and condensation and not by diffusion through the bond.

FIG. 14. Ten-mil copper wire on zinc, 16 hr at 300°C, showing progressive growth of the intermediate phases. ×150.

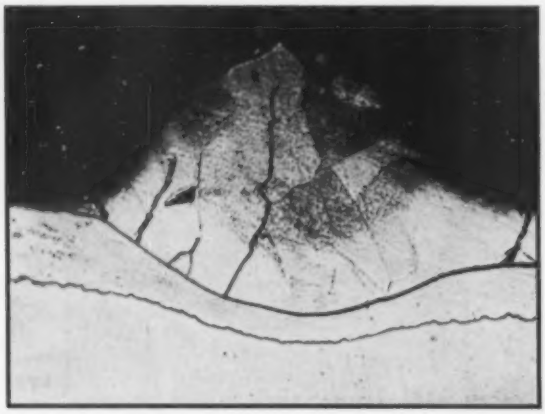


FIG. 15. Five-mil zinc wire on copper, 15½ hr at 384°C. The zinc has mostly evaporated and deposited on the surface of the copper, forming a layer of γ-brass. The core of the original zinc wire has been converted to the dark etching ϵ-phase, ×375.

DISCUSSION

There are several possible explanations of the nonsymmetrical bonding in the bimetal diffusion experiments outlined above. It should be noted that the "digging" always occurs in the metal with the lower melting point. This metal in each case would generally be expected to have lower surface free energy, a higher rate of self-diffusion, and higher vapor pressure than the other metal forming the diffusion couple. A study of the process when both metals are the same has shown that it is best explained by volume diffusion resulting from a vacancy concentration gradient set up at the highly curved junction between the wire and the plate.[3] A straightforward extension of this argument to the case of dissimilar metals would explain the "digging" as the result of faster self-diffusion in the lower melting point metal. That is, atoms flow into the highly curved neck region from the adjacent surfaces of lower curvature and so the latter surface is "eaten away" more rapidly in the metal with the higher diffusion rate. According to this explanation, two grooves of the same dimensions should be observed in the bonding of similar metals. The photomicrograph in Fig. 1 shows this to be the case.

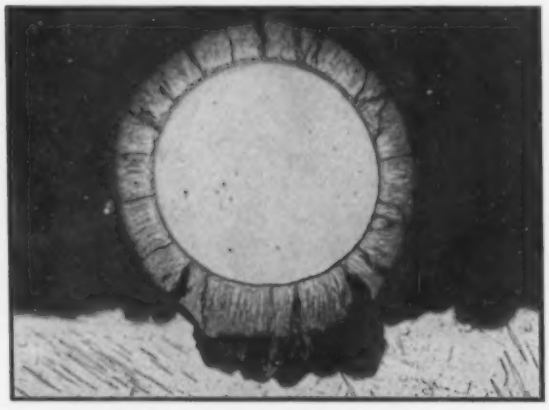


FIG. 16. Five-mil Ni wire on zinc, 2 hr at 370°C. Note the irregular groove on the zinc surface and the columnar alloy layer forming on the Ni wire by condensation of zinc which has evaporated from the zinc base, ×262.5.

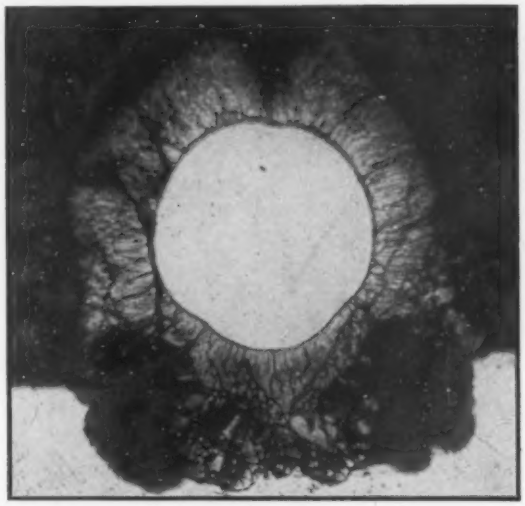


FIG. 17. Five-mil Ni wire on zinc, 4 hr at 370°C. The rim of the widened groove in the zinc is almost in contact with the alloy layer forming on the nickel wire, and this layer in turn is becoming more irregular, ×262.5.

It is also possible to explain the geometry of the interface in these bimetal couples on the basis of surface diffusion, by assuming that the metal of lower surface energy is endeavoring to surround the metal of higher surface energy. This would lead to a "digging" in the manner noted in the examples given above.

It is certainly apparent that evaporation and con-

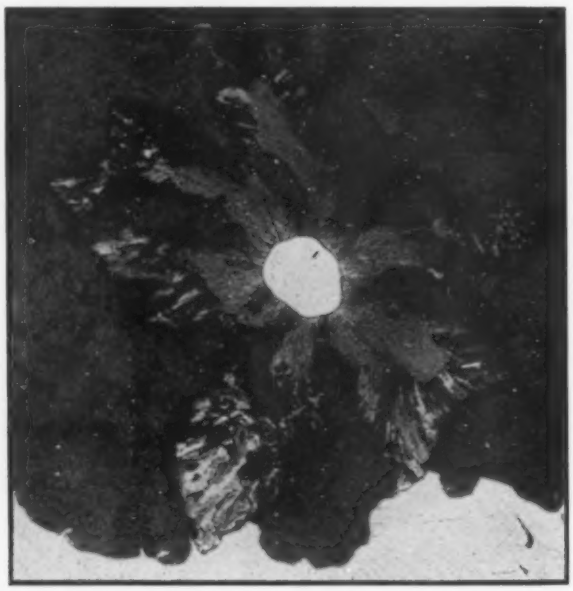


FIG. 18. Same as Fig. 17 except this specimen was heated 16 hr at 370°C. The alloy layer on the nickel wire has now become extremely irregular, ×150.

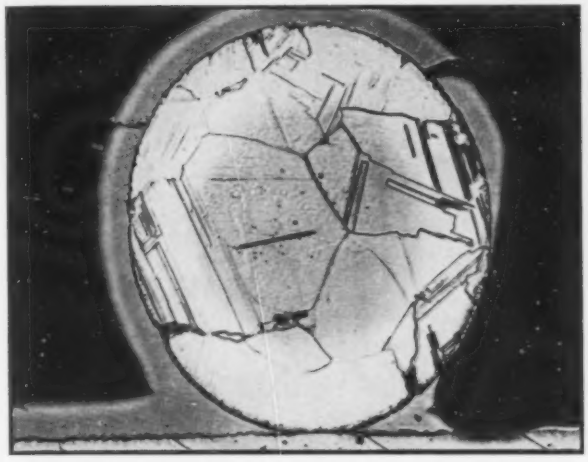


FIG. 19. Ten-mil copper wire on 70/30 brass heated 1 hr at 788°C. Zinc has evaporated from the brass, condensed on the copper wire, and subsequently diffused inward. The bond between the copper and the brass is very poor, ×187.5.

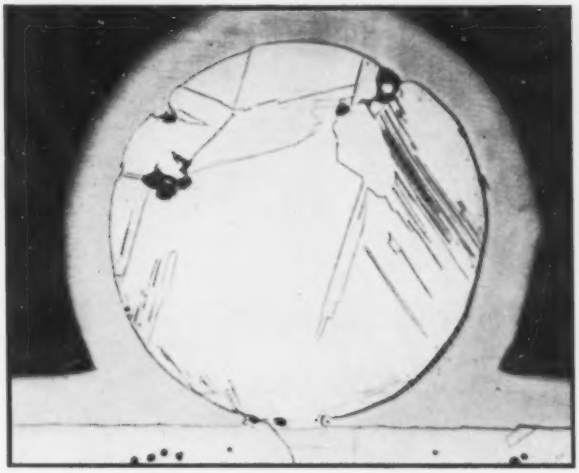


FIG. 20. Ten-mil copper wire on 70/30 brass heated 16 hr at 788°C. The copper wire has been converted entirely to α-brass, and the junction formed is similar to that between copper and copper (Fig. 1), ×187.5.

densation are responsible for much of the change that occurs in the brass and zinc systems where the vapor pressure of zinc is relatively high, and they may also be responsible for some of the changes in the other cases. The fact that the "digging" observed in the copper-nickel diffusion couples was less pronounced in argon than in hydrogen seems to support this. However, it is difficult to see why evaporation should occur preferentially from a region immediately adjacent to the junction. We may also conjecture that the effects described above are due to an unequal flow of the two different kinds of atoms past the original contact, and might then be correlated with the boundary movements observed by Smigelsgas and Kirkendall.[4] Until more work has been done, we cannot be sure of the reasons for the geometrical changes in these diffusion couples. They are important, however, in that the effects of such phenomena are probably manifest in many practical metallurgical operations. The anomalous swelling which accompanies the sintering of powder compacts composed of two different metals (such as Cu-Zn, Fe-Si, Cu-Sn, etc.) is very likely caused by the same factors which account for the interfaces observed in the present diffusion experiments. Powder compacts of two metals often look as if one metal had adsorbed the other, leaving voids. Such effects might also be responsible for the microcracking observed during the homogenization of some castings.

[4] A. Smigelsgas and E. Kirkendall, AIME **171**, 130 (1947).

Scattering and Absorption of Gamma-Rays

M. S. PLESSET*
California Institute of Technology, Pasadena, California

AND

S. T. COHEN
The Rand Corporation, Santa Monica, California

(Received September 27, 1950)

A formulation is presented of the scattering and absorption of gamma-rays in different materials. The range of gamma-ray energies considered is from 1 to 10 mc^2. Results are given for the transmission of gamma-rays through air and lead.

INTRODUCTION

THE release of nuclear energy on a practical scale involves, in general, the emission of large numbers of gamma-rays. The problem of concern in this paper may be stated as follows: for a source of given geometry, intensity, and energy spectrum of the emitted gamma-rays, what are the intensities and energy distributions of these radiations at varying distances from the source? When the intensities and energy distributions are known as functions of distance from the source, the resulting physical and biological effects of the radiations may be determined. Such information is the basis for the determination of the radiological hazards of these radiations and also makes possible the calculation of shielding requirements.

There are three processes which take place as gamma-rays pass through matter: photoelectric effect, pair production, and Compton scattering.

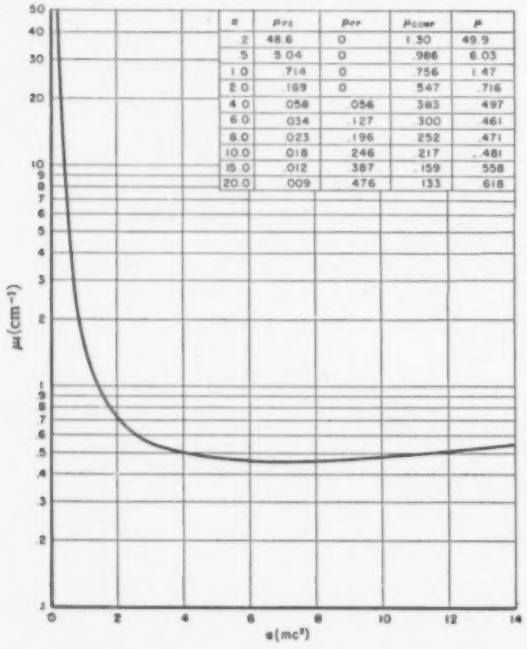


FIG. 1. Total absorption coefficient vs gamma-ray energy, Pb_{82}(lead). [For μ (ordinate and heading of last column) read μ_{81}.]

* Consultant to Rand Corporation.

Photoelectric Effect

The gamma-ray, or photon, is absorbed by an atom of the material with the ejection of one of the bound electrons in the atom. This process has a probability which increases rapidly with atomic number. The probability of this process also increases as the energy of the photon decreases down to the threshold energy characteristic of the atom.

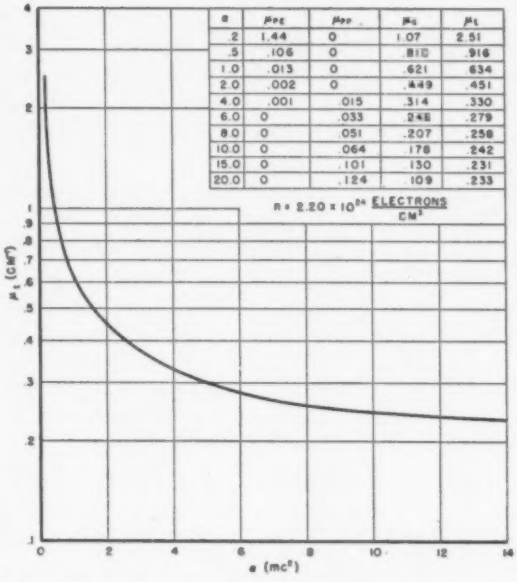
Pair Production

The photon is annihilated in the nuclear field with the creation of an electron and positron pair. The threshold, or minimum, energy of the gamma-ray for this process is the rest energy of the electron pair, 2 mc^2, where m is the rest mass of the electron and c is the velocity of light. The probability of this absorption process increases with the photon energy and with atomic number.

Compton Scattering

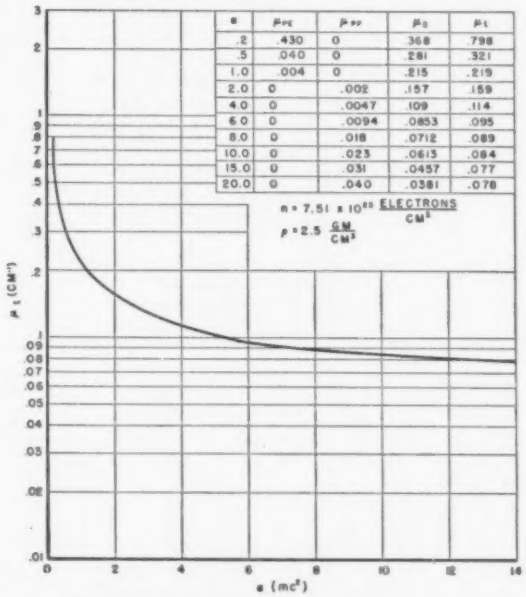
The photon is inelastically scattered by a collision with an electron in the atom. The scattered photon has diminished momentum and energy. The probability of this scattering process decreases with increasing photon energy and is proportional to the number of atomic electrons.

For a quantitative discussion of these effects, it is necessary to define the terms which are customarily used. The probability for gamma-ray absorption or scattering is given in terms of a cross section. For example, the cross section (per atom) for photoelectric absorption, σ_{PE}, is defined as the probable number of photoelectric absorptions divided by the number of incident photons and divided by the number of atoms per cm² of the medium. The atomic cross section, σ_{PP}, for pair production is defined in a similar manner. For Compton scattering, the cross section (per electron) is defined as the probable number of inelastic scatterings divided by the number of incident photons and divided by the number of electrons per cm² of the medium. The photoelectric absorption coefficient is defined as the product of the atomic cross section in cm² by the number of atoms per cm³ and has the dimension cm^{-1}. It is seen, from the definition of the atomic cross section, that the photoelectric absorption coefficient gives the probable



α	μ_{PE}	μ_{PP}	μ_C	μ_t
.2	1.44	0	1.07	2.51
.5	.106	0	.810	.916
1.0	.013	0	.621	.634
2.0	.002	0	.449	.451
4.0	.001	.015	.314	.330
6.0	0	.033	.246	.279
8.0	0	.051	.207	.258
10.0	0	.064	.178	.242
15.0	0	.101	.130	.231
20.0	0	.124	.109	.233

FIG. 2. Total absorption coefficient *vs* gamma-ray energy, Fe_{26}(iron).



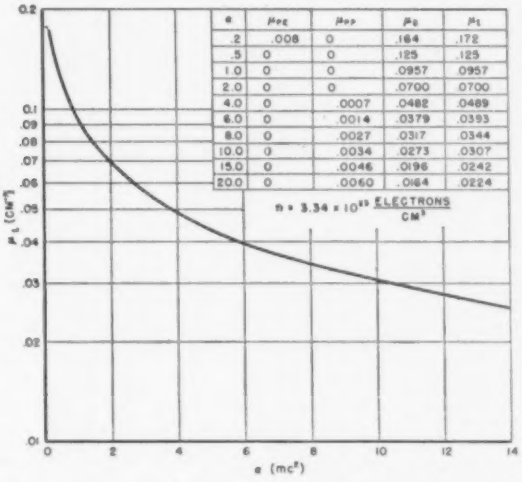
α	μ_{PE}	μ_{PP}	μ_C	μ_t
.2	.430	0	.368	.798
.5	.040	0	.281	.321
1.0	.004	0	.215	.219
2.0	0	.002	.157	.159
4.0	0	.0047	.109	.114
6.0	0	.0094	.0853	.095
8.0	0	.018	.0712	.089
10.0	0	.023	.0613	.084
15.0	0	.031	.0457	.077
20.0	0	.040	.0381	.078

FIG. 3. Total absorption coefficient *vs* gamma-ray energy concrete (SiO_2).

number of photoelectric absorptions per incident photon and per cm of path. The mean free path is defined as the reciprocal of the absorption coefficient. The same definitions apply to the pair production absorption coefficient and the mean free path for this process.

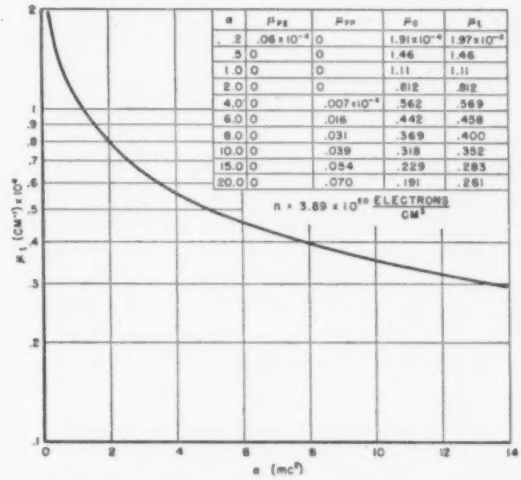
The scattering coefficient for Compton scattering is defined, similarly, as the product of the scattering cross section per electron in cm^2 by the number of electrons per cm^3; and the mean free path is the reciprocal of the scattering coefficient and represents the probable distance traveled by the photon between scatterings.

One may define the total "absorption" coefficient for



α	μ_{PE}	μ_{PP}	μ_C	μ_t
.2	.008	0	.164	.172
.5	0	0	.125	.125
1.0	0	0	.0957	.0957
2.0	0	0	.0700	.0700
4.0	0	.0007	.0482	.0489
6.0	0	.0014	.0379	.0393
8.0	0	.0027	.0317	.0344
10.0	0	.0034	.0273	.0307
15.0	0	.0046	.0196	.0242
20.0	0	.0060	.0164	.0224

FIG. 4. Total absorption coefficient *vs* gamma-ray energy, H_2O.



α	μ_{PE}	μ_{PP}	μ_C	μ_t
.2	$.06\times10^{-4}$	0	1.91×10^{-4}	1.97×10^{-4}
.5	0	0	1.46	1.46
1.0	0	0	1.11	1.11
2.0	0	0	.812	.812
4.0	0	$.007\times10^{-4}$	.562	.569
6.0	0	.016	.442	.458
8.0	0	.031	.369	.400
10.0	0	.039	.318	.352
15.0	0	.054	.229	.283
20.0	0	.070	.191	.261

FIG. 5. Total absorption coefficient *vs* gamma-ray energy, (Air).

gamma-rays as the sum of the photoelectric absorption coefficient, the pair production absorption coefficient, and the Compton scattering coefficient. The total absorption coefficient gives the probability per cm of a removal of a gamma-ray from the primary beam, either by absorption or by scattering. Also, the total mean free path for all three processes is defined as the reciprocal of the total absorption coefficient and gives the probable distance traveled in which any one of the three processes may occur.

Let μ_{PE} be the photoelectric absorption coefficient, μ_{PP} the pair production absorption coefficient, μ_C the Compton scattering coefficient, and $\mu_t = \mu_{PE} + \mu_{PP} + \mu_C$ the total absorption coefficient. The values of these coefficients for lead, iron, concrete, water, and air are tabulated in Figs. 1 through 5 for various values of the photon energy, α. The photon energy, α, is expressed in units of mc^2, the rest energy of the electron, which is 0.51 Mev. In these figures, μ_t is also graphed as a function of α. In the photon energy range of present interest, μ_{PE} and μ_{PP} are very small compared with μ_C for sub-

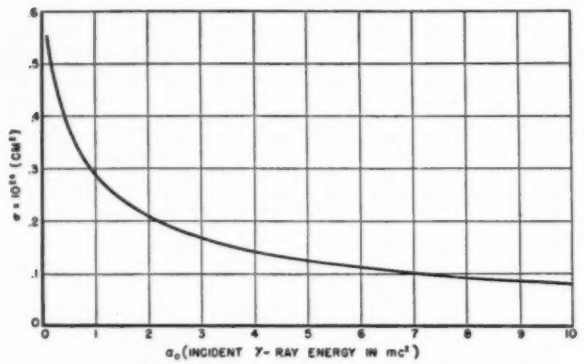


FIG. 6. Total Klein-Nishina cross section per electron *vs* incident gamma-ray energy.

stances such as concrete, water, and air. These substances contain only elements with low atomic numbers Z. Since μ_{PE} is essentially proportional to Z^5, and μ_{PP} is essentially proportional to Z^2, it is evident why these absorption effects become important in the elements of high atomic number.

SCATTERING OF GAMMA-RAYS

As has been pointed out, the scattering process for gamma-rays is Compton scattering. Compton scattering is an inelastic collision between a photon and an electron. If the incident photon of energy $\alpha_0 mc^2$ is deflected through the angle θ after collision with an electron initially at rest, the energy of the scattered photon, $\alpha_1 mc^2$, is determined from the laws of conservation of energy and momentum to be

$$\alpha_1=\frac{\alpha_0}{1+\alpha_0(1-\cos\theta)}. \tag{1}$$

The scattering cross section per electron, σ, for this process is given by the Klein-Nishina formula

$$\sigma=2\pi r_0^2\left\{\frac{1+\alpha_0}{\alpha_0^2}\left[\frac{2(1+\alpha_0)}{1+2\alpha_0}-\frac{1}{\alpha_0}\ln(1+2\alpha_0)\right]+\frac{1}{2\alpha_0}\ln(1+2\alpha_0)-\frac{1+3\alpha_0}{(1+2\alpha_0)^2}\right\}, \tag{2}$$

where r_0 is the classical radius of the electron ($r_0=e^2/mc^2=2.82\times10^{-13}$ cm), and α_0 is the incident photon energy in units mc^2. In Fig. 6, σ is plotted as a function of α_0. The differential cross section for scattering into the element of solid angle $d\Omega=2\pi\sin\theta d\theta$ at θ is

$$\frac{d\sigma}{d\Omega}=\frac{r_0^2}{2}\cdot\frac{1}{[1+\alpha_0(1-\cos\theta)]^2}\times\left\{1+\cos^2\theta+\frac{[\alpha_0(1-\cos\theta)]^2}{1+\alpha_0(1-\cos\theta)}\right\}. \tag{3}$$

The scattering cross section per unit solid angle, $d\sigma/d\Omega$, is shown in Fig. 7 as a function of the angle of scatter θ for various values of the incident energy α_0.

In the determination of the intensities of gamma-radiation as a function of the distance from the source, it is convenient to consider distances in two ranges.

1. Distances small compared with the mean free path.
2. Distances of the order of a mean free path and greater.

For distances small compared with the mean free path, the intensity is accurately given by the transmitted unscattered photons and the singly scattered photons. For distances of the order of a mean free path and greater, the contributions from the photons scattered twice, three times, etc., become increasingly important.

GAMMA-RAY INTENSITIES AT DISTANCES FROM THE SOURCE SMALL COMPARED WITH THE MEAN FREE PATH

The basic scattering problem is the calculation of the intensity at a point a given distance a from a monochromatic (or monoenergetic) point source emitting gamma-rays isotropically. Let N_0 be the number of photons emitted from the source with energy α_0. The number of photons per cm², which arrive at a receiver at distance a from the source, without suffering any collisions, is†

$$I^0=\frac{N_0}{4\pi a^2}\exp[-\mu_t(\alpha_0)a]. \tag{4}$$

The energy received per cm² in units of mc^2 is

$$E^0=\alpha_0 I^0=\frac{\alpha_0 N_0}{4\pi a^2}\exp[-\mu_t(\alpha_0)a]. \tag{5}$$

For distances $a\ll\lambda_t$, the only other contribution of significance comes from singly scattered photons. This intensity is readily computed as follows. The probable

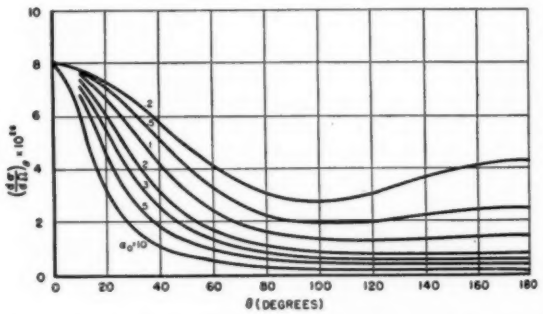


FIG. 7. Klein-Nishina angular distribution scattering cross section per unit solid angle *vs* angle of scatter.

† The exponential factor is easily understood as follows. In terms of the mean free path $\lambda_t=1/\mu_t$, the probability of a collision in the distance dx is dx/λ_t. If $f(x)$ is the probability that the particle will go a distance x without a collision, the probability that the particle will go a distance $x+dx$ without a collision is $f(x+dx)$, and

$$f(x+dx)=f(x)+\frac{df}{dx}dx=f(x)\left[1-\frac{dx}{\lambda_t}\right],$$

or

$$df/dx=-f(x)/\lambda_t=-\mu_t f(x).$$

Thus,

$$f(x)=e^{-\mu_t x}.$$

number of gamma-rays per cm² which reach the element of volume at P (see Fig. 8), is

$$\frac{N_0}{4\pi r_1^2}\exp[-\mu_t(\alpha_0)r_1];$$

the probability of scattering in the element of volume at P into the element of solid angle $d\Omega$ about θ is

$$n\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta} dv_p d\Omega,$$

where n is the number of electrons per cm³, and $dv_p = 2\pi r_2^2 dr_2 \sin\theta d\theta$ is the element of volume at P; and $d\Omega = \Delta S/r_2^2$, where ΔS is the element of area at R as seen from P. The probability that a scattered photon goes from P to R without a collision is $\exp[-\mu_t(\alpha_1)r_2]$, where α_1 is given in terms of α_0 by Eq. (1).

Thus, the number of photons per cm² singly scattered from S to R is

$$d^2I^{(1)} = \frac{N_0}{4\pi r_1^2}\{\exp[-\mu_t(\alpha_0)r_1]\}n\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta} \cdot\frac{2\pi r_2^2 \sin\psi d\psi dr_2}{r_2^2}\exp[-\mu_t(\alpha_1)r_2];$$

$$d^2I^{(1)} = \frac{N_0 n}{2r_1^2}\{\exp[-\mu_t(\alpha_0)r_1-\mu_t(\alpha_0)r_2]\}\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta} \times\sin\psi d\psi dr_2. \quad (6)$$

Further, the singly-scattered energy received per cm² (in units mc^2) is just

$$d^2E^{(1)} = \alpha_1 d^2I^{(1)}. \quad (7)$$

One may express r_1 and r_2 in terms of θ and ψ, so that Eqs. (6) and (7) become

$$I^{(1)}(a) = \frac{N_0 n}{2a}\int_0^\pi d\psi\int_\psi^\pi d\theta\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta} \cdot\exp\left\{-\left[\mu_t(\alpha_0)\frac{a\sin\psi}{\sin\theta}+\mu_t(\alpha_1)\frac{a\sin(\theta-\psi)}{\sin\theta}\right]\right\}; \quad (8)$$

$$E^{(1)}(a) = \frac{N_0 n}{2a}\int_0^\pi d\psi\int_\psi^\pi d\theta\alpha_1\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta} \cdot\exp\left\{-\left[\mu_t(\alpha_0)\frac{a\sin\psi}{\sin\theta}+\mu_t(\alpha_1)\frac{a\sin(\theta-\psi)}{\sin\theta}\right]\right\}. \quad (9)$$

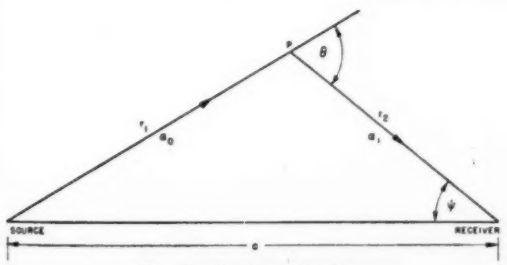


FIG. 8. Single scattering.

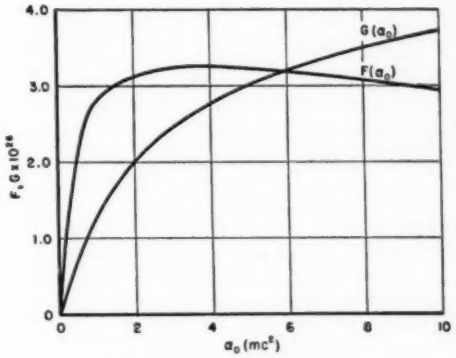


FIG. 9. $F(\alpha_0)$, $G(\alpha_0)$ vs α_0.

The integrals in Eqs. (8) and (9) may be greatly simplified by an approximation which is quite good for energetic gamma-rays, say of energies 2 mc^2 or greater. It is evident (see Fig. 7) that large angle scattering is much less probable at high energies than small angle scattering; further, the exponential factors most heavily weight small values of r_1 and r_2. These conditions mean also that α_1 is near α_0 in the important range of integration. Thus, one makes only a slight overestimate of the intensity if one assumes that

$$\exp[-\mu_t(\alpha_0)r_1-\mu_t(\alpha_1)r_2] = \exp\left\{-\left[\mu_t(\alpha_0)\frac{a\sin\psi}{\sin\theta}+\mu_t(\alpha_1)\frac{a\sin(\theta-\psi)}{\sin\theta}\right]\right\} \approx\exp[-\mu_t(\alpha_0)a]. \quad (10)$$

With this approximation

$$I^{(1)}(a) = \frac{N_0 n}{2a}\{\exp[-\mu_t(\alpha_0)a]\}\int_0^\pi \theta\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta} d\theta; \quad (11)$$

$$E^{(1)}(a) = \frac{N_0 n}{2a}\{\exp[-\mu_t(\alpha_0)a]\}\int_0^\pi a_1\theta\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta} d\theta. \quad (12)$$

The integrals occurring in Eqs. (11) and (12) are functions only of α_0, so that

$$I^{(1)}(a) = \frac{N_0 n}{2a}\{\exp[-\mu_t(\alpha_0)a]\}F(\alpha_0); \quad (13)$$

$$E^{(1)}(a) = \frac{N_0 n}{2a}\{\exp[-\mu_t(\alpha_0)a]\}G(\alpha_0). \quad (14)$$

It should be noted that, to a good approximation, $I^{(1)}(a)$ and $E^{(1)}(a)$ vary with a as $\exp[-\mu_t(\alpha_0)a]/a$, while $I^{(0)}(a)$ and $E^{(0)}(a)$ vary as $\exp[-\mu_t(\alpha_0)a]/a^2$. $F(\alpha_0)$ and $G(\alpha_0)$ are given as functions of α_0 in Fig. 9.

It is customary to express the radiological effects of gamma-rays in terms of the ionization they produce. The conventional unit of this ionization is the roentgen. The roentgen is here taken to be an ionization density

of one electrostatic unit of charge (1 esu) per cm³ of standard air as measured in an ionization chamber with negligible wall thickness. The conversion factor relating gamma-ray intensity with the ionization produced in air is readily found by calculation of the energy transfer to the electrons in air. The energy transfer per cm³ of air is the product of the gamma-ray energy intensity (the energy in units of mc^2 per cm²) by an appropriate absorption coefficient μ_A (cm^{-1}). At the point where it is desired to determine the ionization, let α be the energy of the gamma-rays, and let α' be the energy after a scatter. The energy transfer coefficient, μ_A, is found by expressing the differential cross section [Eq. (3)] in terms of α and α' instead of the initial energy α and the angle of scatter θ. When the differential cross section is so expressed, it will be denoted by $d\sigma/d\alpha'$. Then,

$$\mu_A(\alpha)=\frac{n}{\alpha}\int_{\alpha/(1+2\alpha)}^{\alpha}(\alpha-\alpha')\frac{d\sigma}{d\alpha'}d\alpha', \tag{15}$$

where n is the number of electrons per cm³ of standard air. μ_A is shown as a function of α in Fig. 10. The average energy required to produce an ion pair in air will be taken as 66×10^{-6} mc^2, and the ion charge will be taken as 4.80×10^{-10} esu, so that the ion charge produced in air per cm³ by the energy transfer per cm³, $E\mu_A$, is

$$r\left(\frac{\text{esu}}{\text{cm}^3}\right)=\frac{4.80\times10^{-10}}{66\times10^{-6}}E\mu_A$$
$$=0.727\times10^{-5}\,E\mu_A \text{ roentgens}, \tag{16}$$

where E is the gamma-ray energy intensity (the energy in units of mc^2 per cm²).

The ionization in roentgens produced by the unscattered gamma-rays, from a point source emitting N_0 photons of energy α_0, at a point receiver distant a from the source, is from Eq. (5)

$$r^0(a)=0.727\times10^{-5}\mu_A(\alpha_0)\frac{\alpha_0N_0}{4\pi a^2}\exp[-\mu_t(\alpha_0)a]. \tag{17}$$

The ionization in roentgens produced by singly scattered gamma-rays, from a point source emitting N_0 photons of energy α_0, at a point receiver distant a from the source, is approximately from Eq. (12)

$$r^{(1)}(a)=0.727\times10^{-5}\frac{N_0n}{2a}\{\exp[-\mu_t(\alpha_0)a]\}$$
$$\times\int_0^{\pi}\mu_A(\alpha_1)\alpha_1\theta\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta}d\theta. \tag{18}$$

Equation (18) may be further approximated by replacing the energy spectrum of the scattered gamma-rays by a mean energy $\bar{\alpha}_1$, where

$$\bar{\alpha}_1=\frac{E^{(1)}}{I^{(1)}}=\frac{G(\alpha_0)}{F(\alpha_0)}. \tag{19}$$

Then, an approximate value for the ionization in roentgens of the singly scattered radiation is [Eq. (14)],

$$r^{(1)}(a)\approx0.727\times10^{-5}E^{(1)}(a)\mu_A(\bar{\alpha}_1)$$
$$=0.727\times10^{-5}\mu_A(\bar{\alpha}_1)\frac{N_0n}{2a}G(\alpha_0)$$
$$\times\exp[-\mu_t(\alpha_0)a]. \tag{20}$$

Equations (17) and (20) determine the total ionization in roentgens from a point source when the distance a is much less than the mean free path. The function $[r^{(0)}(a)+r^{(1)}(a)]/N_0$, that is, the roentgens received per emitted gamma-ray, is plotted in Fig. 11 for various values of the gamma-ray energy α_0 when the medium is air.

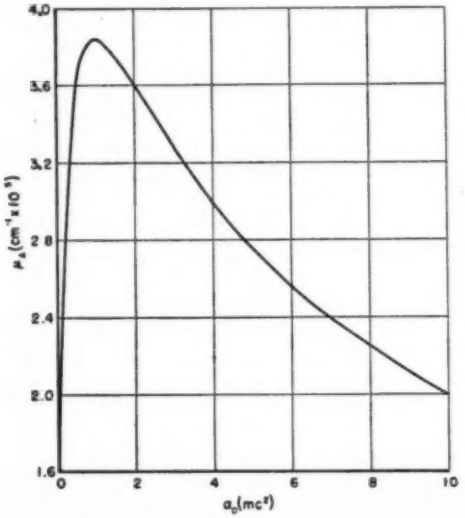


FIG. 10. Absorption coefficient for energy transfer by gamma-rays to air (STP).

GAMMA-RAY INTENSITIES AT DISTANCES FROM THE SOURCE COMPARABLE WITH THE MEAN FREE PATH AND GREATER

For distances of the order of a mean free path or greater, the contributions of the gamma-rays which have been scattered twice, three times, etc., become of increasing importance. The determination of these contributions is a problem of numerical complexity which increases rapidly with the order of scatter. Approximate calculations of the transmission of gamma-rays with multiple scatterings are available;‡ an important feature of these approximations is an essential simplification of the statistics of the sequence of scatterings; e.g., Hirschfelder's approximation consists in replacing the angular scattering distribution by some fixed mean angle of scatter. This approximation overestimates the trans-

‡ Hirschfelder, Magee, and Hull, Phys. Rev. 78, 852 (1948).

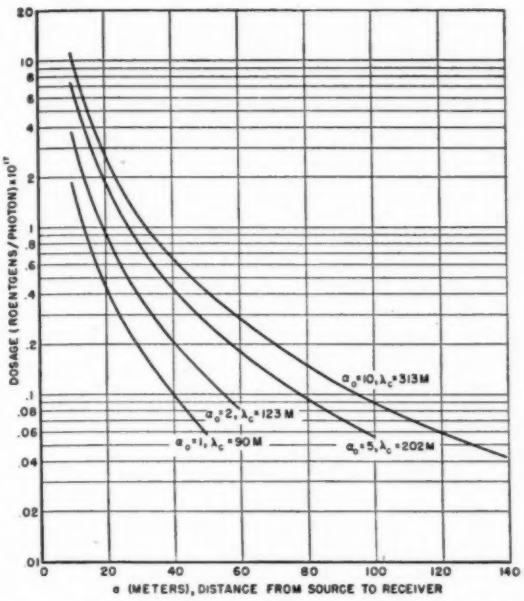


FIG. 11. Dosage in air (STP) per emitted photon from a point source to a point receiver.

mission of the radiation. A more accurate analysis is being made by several investigators, including Mr. H. Kahn and Dr. G. Peebles at RAND Corporation. These studies have been carried far enough so that the total transmitted gamma-ray intensity may be expressed in terms of the incident intensity by means of semi-empirical factors.

The qualitative behavior of the radiation, in the energy range 1 to 10 mc^2 considered here, which has traveled through a thickness of the medium corresponding to many mean free paths may be readily understood. The contribution to the energy intensity, at large distances, of photons which have been scattered through one or more large angles will be small: first, because any large angular deflection requires a considerable reduction in the photon energy; and secondly, because not only does the total absorption coefficient increase, but the probability of another large angle scatter is also increased. One, therefore, expects that the transmitted energy spectrum at large distances will have a mean energy not much less than the initial energy. An important contribution to the transmitted energy intensity will be from photons scattered through relatively small angles.

If $E^{(0)}(a)$ is the unscattered transmitted energy intensity, the total transmitted energy intensity may be written

$$E^t(a)=B[\alpha_0, a\mu_t(\alpha_0)]E^{(0)}(a), \tag{21}$$

and for a point source emitting N_0 gamma-rays isotropically

$$E^t(a)=B[\alpha_0, a\mu_t(\alpha_0)]\frac{N_0\alpha_0}{4\pi a^2}\exp[-\mu_t(\alpha_0)a]. \tag{22}$$

Since the mean transmitted energy is near α_0, one has approximately for the ionization in roentgens

$$r^t(a)\approx 0.727\times10^{-5}\mu_A(\alpha_0)E^t(a)$$
$$=0.727\times10^{-5}\mu_A(\alpha_0)B[\alpha_0, a\mu_t(\alpha_0)]\frac{N_0\alpha_0}{4\pi a^2}$$
$$\times\exp[-\mu_t(\alpha_0)a]. \tag{23}$$

For light elements in which the absorption coefficient $\mu_t(\alpha_0)$ comes primarily from Compton scattering so that $\mu_t(\alpha_0)\approx\mu_C(\alpha_0)$, the factor $B[\alpha_0, a\mu_t(\alpha_0)]$ will vary only slightly from element to element. For heavy elements the factor $B[\alpha_0, a\mu_t(\alpha_0)]$ will also vary with atomic number Z. Transmission in heavy elements will be discussed under "Gamma-Ray Shielding."

Experimental observations have been made in air of the transmission of gamma-rays for which $\alpha_0=3$. These observations show that B becomes a slowly varying factor for large values of $a\mu_C(\alpha_0)$ for $\alpha_0=3$. This is evident from Fig. 12, in which the logarithm of the product of a^2 by the observed ionization is plotted against a. For larger values of α_0, one would expect the factor $B[\alpha_0, a\mu_C(\alpha_0)]$ to increase relatively slowly with increasing α_0.

In Fig. 12, the calculated values of $a^2r^0(a)$ have been drawn for reference; the calculated value of $a^2[r^0(a)+r^{(1)}(a)]$ is also shown.

TRANSMISSION THROUGH A VARIABLE DENSITY MEDIUM

A special problem of importance is the transmission of gamma-rays through an inhomogeneous medium. The

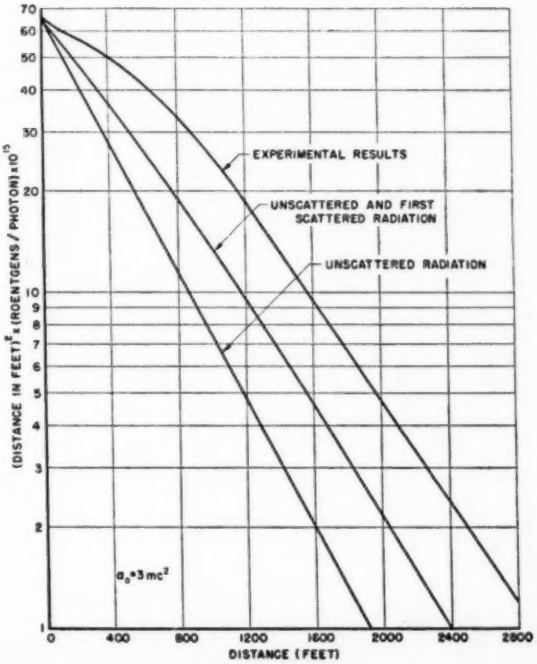


FIG. 12. Dosage in air per emitted photon from a point source to a point receiver.

effect of the inhomogeneity of the medium is taken into account by revising the exponential attenuation factor $\exp[-\mu_t(\alpha_0)r]$ to become $\exp[-\int_0^r \mu_t(\alpha_0, r)dr]$. Consider, as an example, the transmission of gamma-rays through a variable atmosphere. For air, where $\mu_t(\alpha_0) \approx \mu_C(\alpha_0)$, the effect of the variable atmosphere is taken into account by replacing $\mu_C(\alpha_0)a$ with

$$\int_0^a \mu_C(\alpha_0, z)dz.$$

Thus,

$$E^1(a) = B[\alpha_0, a\bar{\mu}_C(\alpha_0)]\frac{N_0\alpha_0}{4\pi a^2} \times \exp\left[-\int_0^a \mu_C(\alpha_0, z)dz\right]. \quad (24)$$

The factor B is a slowly varying function of a for the large values of $a\mu_C(\alpha_0)$ of present interest. One has

$$\mu_C(\alpha_0, z) = \frac{A}{M} Z\sigma(\alpha_0)\rho(z),$$

where A is Avogadro's number, M is the molecular weight of air, Z is the effective atomic number of air, and $\rho(z)$ is the atmospheric density. Thus,

$$\int_0^a \mu_C(\alpha_0, z)dz = \frac{A}{M} Z\sigma(\alpha_0)\int_0^a \rho(z)dz.$$

This density integral may be evaluated.

For example, one may approximate the atmosphere in the troposphere by a dry adiabatic atmosphere in which

$$p = p_0 e^{-(gz/RT)},$$

and

$$T = T_0 - \beta z,$$

where g is the acceleration of gravity, R is the gas constant (in ergs per g per °K), T is the temperature in °K, p is pressure, and β is the dry adiabatic lapse rate which

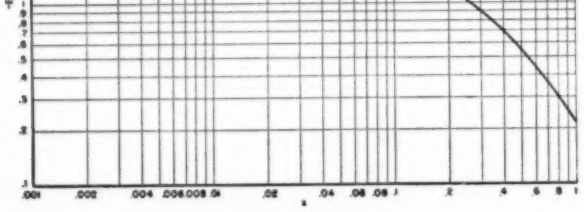


FIG. 13. Exponential integral function.

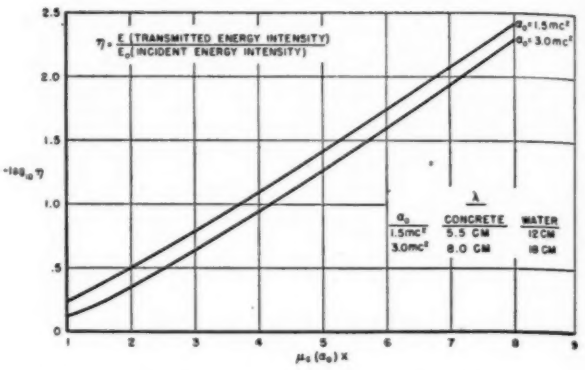


FIG. 14. Energy attenuation of gamma-rays in water and concrete.

is 9.8°K per 1000 meters. Then, one finds directly that

$$\int_0^a \rho(z)dz = e^{g/R\beta}\frac{\rho_0 T_0}{\beta}\left[Ei\left(-\frac{g}{R\beta}\frac{T_0}{T_0-\beta a}\right) - Ei\left(-\frac{g}{R\beta}\right)\right],$$

and finally

$$\int_0^a \mu_C(\alpha_0, z)dz = \frac{A}{M} Z\sigma(\alpha_0)e^{g/R\beta}\frac{\rho_0 T_0}{\beta} \times \left[Ei\left(-\frac{g}{R\beta}\frac{T_0}{T_0-\beta a}\right) - Ei\left(-\frac{g}{R\beta}\right)\right]. \quad (25)$$

A graph of the exponential integral function $-Ei(-x)$ is given in Fig. 13.

GAMMA-RAY SHIELDING

Shielding with Light Elements

In light elements, the total absorption coefficient (in the energy range of present practical interest), is given for the most part by the Compton scattering coefficient. The shielding problem is essentially characterized by considering the attenuation of a plane beam of gamma-rays incident upon a plane slab of the shielding material of thickness X. When the distance from the source is large, the incident gamma-radiation may be taken as a plane beam.

The attenuation of a plane beam of the radiation incident upon the shield may be readily calculated, for thicknesses small compared with the mean free path, by determination of the unscattered and singly-scattered transmitted intensities. If one has I_0 photons per cm² of energy α_0 normally incident upon a plane slab, the unscattered transmitted intensity is

$$I^0(X) = I_0 \exp[-\mu_C(\alpha_0)X], \quad (26)$$

and the transmitted energy intensity is

$$E^0(X) = \alpha_0 I_0 \exp[-\mu_C(\alpha_0)X]. \quad (27)$$

The singly-scattered transmitted intensity $I^{(1)}(X)$ is given by

$$I^{(1)}(X)=2\pi nI_0\int_0^{\pi/2}\sin\theta d\theta \times\exp[-\mu_C(a_1)X/\cos\theta]\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta}\int_0^X dx \times\exp\left\{-x\left[\mu_C(\alpha_0)+\frac{\mu_1(\alpha_1)}{\cos\theta}\right]\right\},$$

$$I^{(1)}(X)=I_0\exp[-\mu_C(\alpha_0)X]\int_0^{\pi/2}2\pi\sin\theta d\theta n\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta} \times\left[\frac{1-\exp\left\{-\frac{X}{\cos\theta}[\mu_C(\alpha_1)-\mu_C(\alpha_0)\cos\theta]\right\}}{\frac{\mu_C(\alpha_1)}{\cos\theta}-\mu_C(\alpha_0)}\right], \quad (28)$$

and the transmitted, singly-scattered energy intensity is

$$E^{(1)}(X)=E_0\exp[-\mu_C(\alpha_0)X] \times\int_0^{\pi/2}\frac{2\pi\sin\theta d\theta}{1+\alpha_0(1-\cos\theta)}n\left(\frac{d\sigma}{d\Omega}\right)_{\alpha_0,\theta} \times\left[\frac{1-\exp\left\{-\frac{X}{\cos\theta}[\mu_C(\alpha_1)-\mu_C(\alpha_0)\cos\theta]\right\}}{\frac{\mu_C(\alpha_1)}{\cos\theta}-\mu_C(\alpha_0)}\right]. \quad (29)$$

The primary interest is in the transmitted energy intensity, since this is directly correlated with ionization produced at a receiver. The ionization produced in roentgens by the transmitted intensity is

$$r=0.727\times10^{-5}[\mu_A(\alpha_0)E^{(0)}(X)+\mu_A(\bar{\alpha}_1)E^{(1)}(X)], \quad (30)$$

where $\bar{\alpha}_1$ is very nearly $E^{(1)}(X)/I^{(1)}(X)$ and may even be replaced by α_0 with only a small loss in accuracy.

For thicknesses of several mean free paths, one has recourse to the semi-empirical factor as discussed under "Gamma-Ray Intensities at Distances from the Source Comparable with the Mean Free Path and Greater." The transmitted energy intensity is

$$E[\mu_C(\alpha_0)X]=E_0\exp[-\mu_C(\alpha_0)X]B[\alpha_0,\mu_C(\alpha_0)X]. \quad (31)$$

Attenuation coefficients for water and concrete are shown in Fig. 14 for $\alpha_0=1.5$ and 3.

TABLE I. Lead.

X (cm)	$B(X)$
0	1.0
2	1.2
4	1.4
6	1.6
8	1.9
10	2.2
12	2.5
14	2.8
16	3.1
18	3.5
20	4.0
22	4.5
24	5.0

Shielding with Heavy Elements

For heavy elements, μ_{PP} and μ_{PE} make important contributions to μ_t. In elements of high atomic number, such as lead, and in the energy range of present concern, the photons which are degraded in energy by Compton scattering have rapidly increasing probability of complete absorption. As a result, the calculation of the attenuation of gamma-rays in such elements is a problem involving the determination of a relatively small number of scatterings.

Numerical calculations have been made of the total energy intensity transmitted through various thicknesses of lead for several values of the incident energy α_0. These calculations indicate the persistence of the "hard core" in the transmitted energy; that is, the energy spectrum of the transmitted beam is heavily weighted toward the incident photon energy α_0. If one writes the total transmitted energy intensity $E(X)$ as

$$E(X)=E_0B(X)\exp[-\mu_t(\alpha_0)X], \quad (32)$$

where E_0 is the incident energy intensity. Then $B(\alpha_0, X)$ has been calculated to have the approximate values given in Table I for $\alpha_0=10$.

Further calculations for different values of the incident energy α_0 show that the ratio of the total transmitted energy intensity to the unscattered energy intensity varies only slightly with the incident photon energy in the range $2\leqslant\alpha_0\leqslant10$ and is, thus, approximately only a function of the thickness X. Thus, one may use Table I to obtain rough values of $B(X)$ for this range of incident gamma-ray energies.

The ionization in roentgens produced by the transmitted energy is, thus, approximately

$$r=0.727\times10^{-5}\mu_A(\alpha_0)E_0B(X)\exp[-\mu_t(\alpha_0)X]. \quad (33)$$

Letters to the Editor

The Elastic Parameters of Beta-Brass

ROBERT A. ARTMAN* AND DONALD O. THOMPSON
Department of Physics, State University of Iowa, Iowa City, Iowa
November 1, 1950

As part of an investigation into some of the physical properties of beta-brass on and near the copper-rich side of the beta-phase boundary, the elastic parameters of beta-brass at 25°C were measured. Inasmuch as the results show a composition dependence of the elastic parameters, and comparison with previous results[1-4] reveals the same phenomenon, it is felt that a report on this effect is worthwhile.

Single crystals of beta-brass 6.3 mm in diameter and from 8 cm to 15 cm in length were grown in graphite molds by the Bridgman method. Five crystals whose composition was shown to be just at the copper-rich side of the phase boundary by chemical analysis and micrographical examination were used in the determination of the constants. The average composition of the set is 55.12 atomic percent copper, 44.88 atomic percent zinc (difference). Orientations were determined by back reflection diffraction pictures. Young's modulus, E, was measured dynamically using a composite piezoelectric oscillator,[5] and the rigidity modulus, G, was measured statically with an apparatus designed to produce pure axial torque only. The difference in method offers no difficulty since s_{44} is the same whether measured statically or dynamically. A plot of reciprocal Young's moduli against orientation function gives s_{11} directly, and a similar plot of reciprocal rigidity moduli gives s_{44} directly, both as intercepts of the plot. In the first case the slope of the line is $-2s$, and in the second it is $4s$, where $s=s_{11}-s_{12}-\frac{1}{2}s_{44}$. By combination of these, the three elastic parameters, s_{11}, s_{12}, and s_{44} are determined.

The results of this and previous experiments[2,4] are compared in Table I.

TABLE I.* Adiabatic elastic parameters of beta-brass.

	Rinehart[a]	Lazarus[b]†	Authors
Atomic percent copper	52.47	51.74	55.12
$1/E_{100}=S_{11}$(cm²/d)	3.881×10^{-12}	3.517×10^{-12}	4.105×10^{-12}
$-S_{12}$		1.616	1.898
$1/G_{100}=S_{44}$		1.213	1.344
$1/E_{111}$	0.475	0.499	0.551
S	5.109	4.527	5.331

* Good's values (see reference 3) are not included in Table I, since it now seems certain that his value of s_{44} is too low, perhaps because, as pointed out by Lazarus, of the existence of an impure mode involving coupling between torsion and flexure in the composite piezoelectric oscillator method used.
† In transforming Lazarus' results to values of the S's (from C's) his directly measured value of $(C_{11}-C_{12})$ has been used rather than the difference in C_{11} and C_{12} being taken as given in his paper.
[a] See reference 2.
[b] See reference 4.

The difference between Lazarus and the present authors is believed to be due entirely to composition difference. Measurements of other crystals differing in composition from those in Table I show that both their $1/E$ and $1/G$ values agree with values obtained by interpolation between the values of Lazarus and the authors. Interpolation is based on the observation that all of the constants listed by Lazarus and the writers are nearly in the same ratio as the ratio of the ratios of the number of copper atoms to zinc atoms for the two compositions. All the measured values of $1/E$ by Rinehart lie between those of Lazarus and the writers although the line drawn through his points determines a lower value of $1/E_{111}$ than either of the other two sets of data. Also, his values lie nearer those of the writers, while the composition is nearer that of Lazarus. Rinehart's composition is, however, that of the stock material and not of the crystals. A small loss of zinc during growth of the crystals would be entirely sufficient to explain the observed relation of his results to the other sets of data. In addition, the scatter of Webb's results[1] can be reduced considerably by application of the interpolation procedure to the average composition of his crystals.

TABLE II.

	C'	K	A	σ_{100}	σ_{111}
Present	0.0833	1.079	8.933	0.462	0.220
Lazarus	0.0974	1.162	8.465	0.459	0.213

Table II lists interesting quantities which may be studied as a function of composition. C' is the shear modulus $1/[2(S_{11}-S_{12})]=\frac{1}{2}(C_{11}-C_{12})$, K is the bulk modulus $1/[3(S_{11}+2S_{12})]=(C_{11}+2C_{12})/3$, A is the anisotropy ratio $2(S_{11}-S_{12})/S_{44}=C_{44}/C'$, $\sigma_{100}=-S_{12}/S_{11}$ is "Poisson's ratio" for a specimen subjected to a tensile stress along the ⟨100⟩ axis, and $\sigma_{111}=(3S_{12}+S)/(2S-3S_{11})$ is "Poisson's ratio" for a specimen subjected to a tensile stress along ⟨111⟩ axis.

Perhaps the most significant quantity in Table II is C', the value of this quantity decreasing with increasing copper concentration. This is consistent with theory, for, as pointed out by Zener,[6] in a body-centered cubic copper lattice the contribution of the exchange interaction to C' is negative, whereas the contribution of the electrostatic interaction between conduction electrons and ions while positive is independent of the ion structure. Substitution of copper atoms for zinc would decrease the value of C'. The ratio of the two values of C' is nearly the same as the reciprocal of the interpolation ratio. Furthermore, the smaller value of C' for copper rich compositions means decreasing mechanical stability for the crystal structure as might be expected for compositions nearer the alpha-plus-beta-phase boundary.

It is worth pointing out that either value of the anisotropy ratio, A, is considerably lower than that derivable from previous results,[3] and puts beta-brass in a class with Na and K in this respect. Also, either value of σ_{111} is positive, whereas its value derivable from previous results[3] was negative, the only exception in metals to positive values for this quantity.

The preceding data were taken as part of an investigation of Young's modulus and internal friction from room temperature to above the order-disorder transition and direct measurement of the bending produced by axial torque. These results will be reported in a subsequent paper.

* Now at Evansville College, Evansville, Indiana.
[1] W. Webb, Phys. Rev. 55, 297 (1939).
[2] J. S. Rinehart, Phys. Rev. 58, 385 (1940).
[3] W. A. Good, Phys. Rev. 60, 605 (1941).
[4] D. Lazarus, Phys. Rev. 74, 1726 (1948), and Phys. Rev. 76, 545 (1949).
[5] S. L. Quimby, Phys. Rev. 25, 558 (1925).
[6] C. Zener, Phys. Rev. 71, 846 (1947).

The Barium Oxide on Tungsten Cathode Interface

H. P. ROOKSBY AND E. G. STEWARD
Research Laboratories, The General Electric Company, Limited, Wembley, England
November 11, 1950

In a recent letter in these columns, Hensley and Affleck[1] have referred to interface compounds formed under the alkaline earth oxide coating on tungsten base cathodes. They point out that compounds different from the usual tungstates, such as $BaWO_4$, are produced, and they suggest that these compounds are tungstites of the type $BaWO_3$. A structural analogy with members of the perovskite family is drawn, and the x-ray data obtained from interfaces formed with oxide coatings containing strontium as well as barium are interpreted accordingly.

Evidence we have obtained in the course of our investigations on oxide cathodes, and in particular on cathode interface compounds,[2] leads to conclusions different from those put forward by Hensley and Affleck. We believe that their interpretations both

of the constitution and of the crystal structure of the interface compounds on tungsten cathodes are incorrect.

We have succeeded in synthesizing alkaline earth tungstates which are similar in crystal structure to the tungsten interface compounds. The synthesis does not necessarily have to be performed in vacuum, but can be accomplished by firing appropriate mixtures of alkaline earth carbonate and tungstic oxide in air at a sufficiently high temperature. The alkaline earth carbonate and tungstic oxide must be in the molecular proportions 3:1, so that the compounds have a molecular formula R_3WO_6, where R is the alkaline earth metal. In a typical instance, a tungstate of composition Ba_2CaWO_6 can be prepared by firing the appropriate mixture of barium carbonate, calcium carbonate, and tungstic oxide in air at 1400°C. The crystal structure and lattice parameters correspond with those of the interface compound found on tungsten when the coating is a mixture of barium and calcium oxides.

Such alkaline earth tungstates do possess, as Hensley and Affleck found, a crystal structure of cubic or near-cubic symmetry, but the space group is not that of the ideal perovskite structure.[3] As shown subsequently, when all reflections on x-ray powder photographs are taken into account the sequence is consistent with face-centered cubic symmetry, and differs from that characteristic of the perovskite space group. This is in addition to the splitting and multiplicity of high order lines, which indicate a lowering of symmetry from a true cubic arrangement.

Attention was drawn by us to the existence of these alkaline earth tungstates in a letter to *Nature* in 1946.[4] The identification of such compounds as interface reaction materials on tungsten-cored oxide cathodes was also mentioned. For comparison with the data given by Hensley and Affleck, we include here Table I listing lattice constants determined for representative members of the new class.

TABLE I.

Tungstate compound	Approx. structure cell constant, assuming cubic symmetry. (kX)
Ba_3WO_6	8.6_0
Ba_2SrWO_6	8.5_5
Ba_2CaWO_6	8.3_8
Sr_3WO_6	8.2_7
Ca_3WO_6	8.0_9
$BaSrCaWO_6$	8.2_7

A cubic structure cell is assumed, and departures to a greater or lesser degree from the isometric system have been ignored. Consequently, in instances where large deformations of the cubic structure cell are manifest in the x-ray patterns, the values tabulated are very approximate.

In Fig. 1 (a) intensities (visually estimated and uncorrected) are plotted against Bragg angles θ (copper radiation), for all the reflections up to $\theta \sim 40°$, observed in a photograph of Ba_2CaWO_6. This particular tungstate exhibits less distortion from true cubic symmetry than most of the others given in the table and is, therefore, more suitable for comparison purposes. With the hkl values of a face-centered cubic structure the cell dimension is 8.3_8kX.

As shown in Fig. 1 (b), lines may be selected from the above set, giving a pattern similar to that of $CaTiO_3$ (Fig. 1 (c)). Indexing such reflections on a perovskite structure basis, gives a cell dimension of half that obtained above *viz.* 4.1_9kX. In a similar way, the different indexing also explains why the value of the lattice constant of 8.5_5kX listed above for Ba_2SrWO_6 is roughly double the 4.24kX quoted by Hensley and Affleck for their "barium strontium tungstite", 8.2_7kX compared with 4.12kX for their "strontium tungstite", and 8.6_0kX compared with 4.30kX for their "barium tungstite".

The similarity to the perovskite pattern was demonstrated by Hensley and Affleck in the case of their "barium tungstite", but the often weak and intervening lines which provide evidence of the face-centered structure were either not detected, or, in the instance where Hensley and Affleck do refer to "a few other lines . . . sometimes present", misinterpreted.

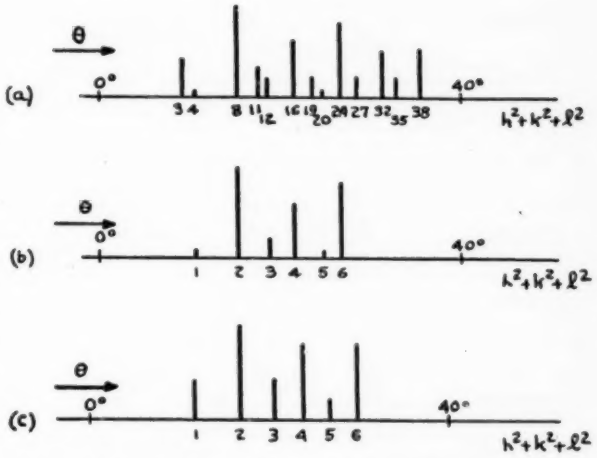


FIG. 1. (a) Intensities *vs* Bragg angles θ for Ba_2CaWO_6; (b) selected reflections from (a) for comparison with (c); (c) intensities *vs* Bragg angles θ for $CaTiO_3$.

The observation that the barium compound could be formed from a mixture of $BaCO_3$ and powdered tungsten on heat treatment in a very high vacuum may be explained in the following manner. For production of the Ba_3WO_6 compound from such a mixture, limited oxidation of the tungsten is first of all required, and such oxidation is a possible consequence of degassing of the barium carbonate. It follows as a result of the limited oxidation that a high concentration of barium oxide in proportion to tungsten oxide exists in the mixture. Thus, Ba_3WO_6 is formed in preference to $BaWO_4$ provided a high reaction temperature, say, 1300°C or more, is reached. We have carried out similar heat treatments, both in vacuum and in an inert atmosphere, of mixtures of barium carbonate and metallic tungsten. The products from equimolecular mixtures often contained much Ba_3WO_6, but they invariably contained free tungsten and unreacted barium oxide as well.

[1] E. B. Hensley and J. H. Affleck, J. Appl. Phys. **21**, 938 (1950).
[2] H. P. Rooksby, Nature **159**, 609 (1948).
[3] I. Naray-Szabo, Mücgyetemi Közlemenyck **1**, 30 (1947).
[4] H. P. Rooksby and E. G. Steward, Nature **157**, 548 (1946).

The Barium Oxide on Tungsten Cathode Interface

EUGENE B. HENSLEY AND JOHN H. AFFLECK
University of Missouri, Columbia, Missouri
December 26, 1950

IN a recent Letter[1] to the Editor in this Journal we presented evidence for believing the interface compound formed in a barium oxide on tungsten cathode had a perovskite-type crystal structure with a formula $BaWO_3$. Rooksby and Steward have called our attention to a letter in *Nature*,[2] in which they identified the compound as Ba_3WO_6 with a face-centered pseudocubic structure having exactly twice the linear dimensions proposed by us. On re-examining our diffraction patterns, we have found among the weak background lines the additional lines predicted by the face-centered structure. We were misled in our conclusions by the fortuitous agreement of the diffraction line intensities with those predicted by the assumed perovskite-like structure. Also, by following Rooksby and Steward's suggestion, we have now succeeded in synthesizing the interface compound by heating in air a mixture of several parts of $BaCO_3$ to each part of WO_3 at 1300°C.

[1] E. B. Hensley and J. H. Affleck, J. Appl. Phys. **21**, 938 (1950).
[2] H. P. Rooksby and E. G. Steward, Nature **157**, 548 (1946).

Water Movement Induced by Compaction of Clastic Sediments

ROBERT A. HOWARD AND CARL E. BLEIL
Department of Physics, University of Oklahoma, Norman, Oklahoma
December 11, 1950

THE motion of water through rock strata is often considered in theories on the migration of petroleum. The pressure gradient required to sustain the velocity of the water could also act to cause motion of oil within the interstices of the rock, and, if sufficiently great, could even overcome opposing capillary forces; that is, forces arising from interfacial and adhesion tensions. A possible origin of moving water is to be found in the compaction of sediments which are laid down in water in an initially poorly-packed state. The accumulation of overburden would then increase the compaction of the buried sediments, and water would be squeezed from them and have to escape to the surface.

If it be assumed that the water flows vertically through the sediments to the surface and that equilibrium compaction is maintained at all times (i.e., viscous forces do not cause a lag in the attainment of compaction for a given amount of overburden), then it is possible to compute the rate of flow of water in the following manner. Let the porosity, which is defined as the percent of pore volume, or, in this case, the volume of water per unit volume of sediment, of the freshly deposited sediments be F_0; and assume that the porosity under an infinite overburden is reduced to zero. The porosity may then be represented functionally by $F=F_0G(z)$, where z is the depth of burial from the surface and is measured positively downward, and $G(z)$ is a function which is always less than or equal to one, and in particular $G(0)=1$ and $G(\infty)=0$.

Let y be the depth at which a particle at z would have been buried had there been no compaction.

If compressibility of the water and solid particles be neglected, then conservation of matter requires that $(1-F)dz=(1-F_0)dy$ or $dz/dy=(1-F_0)/1-F$. Let J be the current (volume of water per unit time) passing upward through a horizontal surface of unit cross section at a depth z. The conservation condition then leads to the expression $dJ=d(Fdz)/dt$, where dJ is the excess of fluid flowing into the volume dz at $z+dz$ over the amount flowing out at z per unit time. Now $d(Fdz)/dt=(dFdz/dt)+Fd(dz/dt)$; and if the rate of deposition of sediment is assumed to be constant, so that $dy/dt=A$, then $dJ=A(1-F_0)dF/(1-F)^2$. Integration of both sides of this expression for the case of infinitely thick sediments yields

$$J=\int_0^J dJ=A(1-F_0)\int_0^F dF/(1-F)^2=AF[(1-F_0)/1-F].$$

This expression yields the curve shown in Fig. 1 for the special

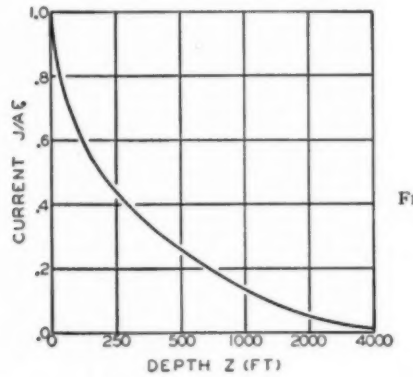


FIG. 1.

case where $F=F_0e^{-z/b}$. The exact shape of the curve depends upon the function F, but the upper and lower values of J are dependent only on the limiting values of F.

The average linear velocity* of a water molecule at depth z is given by $V=J/F=A[(1-F_0)/1-F]$. This function is shown for the special case of $F=F_0e^{-z/b}$ in Fig. 2, the exact shape of the curve

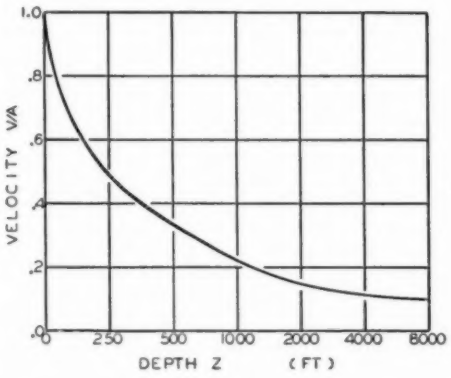


FIG. 2.

again depending upon the form of F. The velocity approaches the constant value $A(1-F_0)$ at great depths, irrespective of the exact manner in which F varies with z, provided that $F\to 0$ as $z\to\infty$.

It is to be emphasized that the velocities computed in this way are essentially *maximum* velocities. For example, if the sediment extends only to a finite depth and is underlain by nonporous basement rock, then the current is reduced throughout by an amount equal to the value of J at this depth computed from the above formula. If the sediments are not all compactible materials, such as shale, but have interbedded layers of only slightly compactible material, such as sandstone, then the flow of water will be reduced. The velocity would also be reduced if the compactible sediments reached a minimum value of porosity greater than zero at great depth. The viscosity of the water, which has been ignored in this treatment, can have only the effect of *decreasing* the rate of flow of the water.

There is the possibility that water pressed from underlying formations will escape, not vertically through the overlying formations, but laterally through a highly permeable layer. In order to be effective, the specific fluid conductivity of such a layer must be so high that its conductivity is greater than that of the overlying formations to vertical flow in spite of the enormously greater area across which flow can occur in the case of vertical movement. Even if fluid does escape by such lateral permeable beds, the vertical flow through the underlying rocks will not exceed the value given by the formulas derived here.

It is felt, therefore, that the formulas derived here should serve as a satisfactory means for estimating maximum water flow velocities to be considered in the evaluation of theories of petroleum migration.

* The average linear speed of a water molecule would be somewhat greater than this average velocity because of the tortuosity of the channels through which it must move.

Impact Flash at High Speed Water Entry

J. H. MCMILLEN, R. L. KRAMER, AND D. E. ALLMAND
U. S. Naval Ordnance Laboratory, White Oak, Silver Spring, Maryland
December 11, 1950

WHILE it is common knowledge that light may be emitted when solid matter is struck by a fast missile, it is not generally known that light is also produced when a fast missile strikes a water surface. This phenomenon is observed in our laboratory when ½-inch steel spheres are shot vertically downward into water with speeds up to 7000 ft/sec.

The flash may be seen visually in a darkened room, and it has been recorded photographically by using an open shutter camera.

At present, this short flash has been used in conjunction with a photo-multiplier tube to trigger the spark in our spark-shadowgraphic equipment. The evidence we have at present shows that the light is emitted close to the surface, certainly within one sphere diameter.

It has been possible to obtain simultaneous records of the impact flash and the water-entry spark shadowgram, by placing the shadowgraphic film in the water within ¼ inch of the missile's path. The shadowgram contains a record of the time-integrated illumination from the impact flash and one of the instantaneous shadowgram of the impact shock wave and missile cavity. The impact flash record in this case is simply one of fogging of the film near the impact point. One of these shadowgrams is reproduced in Fig. 1, where the shock wave has advanced eight sphere

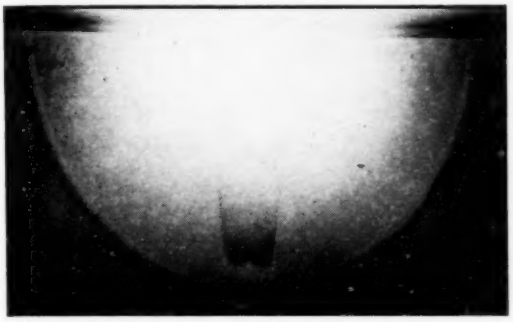


FIG. 1.

diameters into the water. The missile in this case had an impact velocity of 6900 ft/sec. A similar shadowgram is shown in Fig. 2

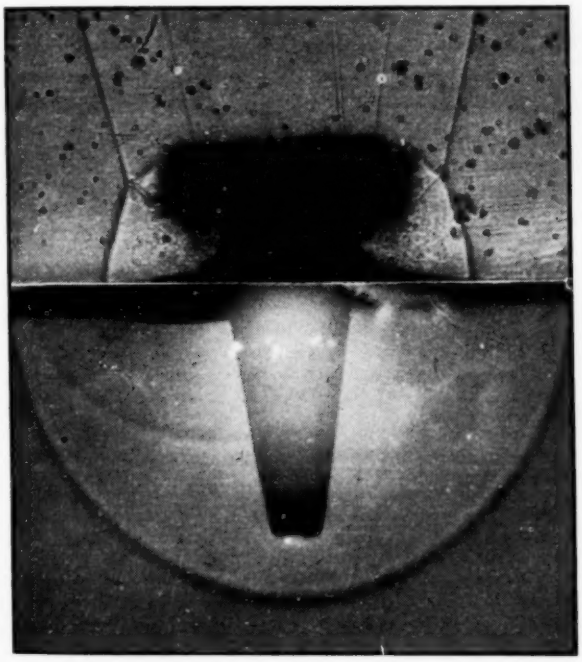


FIG. 2.

except that fogging effects are reduced by the introduction of a thin cardboard light-shield between the impact point and the film. In Fig. 2, the shadowgram was produced on two separate films; the film above the water surface showing the air shocks, and splash was placed 24 sphere diameters farther from the impact point than was the lower water-immersed film. The light is predominantly below the surface and seems to be directed in a downward direction. This directional effect may have been produced by reflections from the water surface, which in the early stages of entry rises up the sides of the sphere. Most shadowgrams show a strong concentration of light at an angle of 45 degrees with the missile's path.

Several explanations for this impact flash have occurred to the authors, but no evidence is at hand to support them. One interpretation supposes that the lateral splash possesses a velocity much greater than that of the sphere and that it is sufficiently high to maintain an ionizing stagnation temperature in the air. However, the predominance of light below the water surface does not support this interpretation. Another interpretation is that the adiabatic compression at impact is sufficient to make either the sphere or the water momentarily self-luminous. However, a calculation of the pressures required to produce these self-luminous temperatures gives higher values than are customarily supposed to exist at impact. It is interesting to note that individual water molecules which acquire the same speed as the missile at impact will have an equivalent temperature of 2800°C, which is ample for self-illumination.

The Crystal Structure of the Sigma-Phase in the Co-Cr System

J. S. KASPER, B. F. DECKER, AND J. R. BELANGER
General Electric Research Laboratory, Schenectady, New York
December 11, 1950

THE technical importance of the sigma-phase, which has been found to occur in many systems, has made it the concern of many investigators. In recent years, much effort has been expended in attempting an indexing of the powder patterns, and various possible cell dimensions have been reported in the literature. In this laboratory, similar efforts have been made over a number of years with the powder patterns of the sigma-phase in the systems of Fe-Cr, Fe-V, Co-Cr, Co-Ni-Cr, and Mo-Mn. Although several possible indexings have been found which were as satisfactory as those in the literature, they were subject to the usual uncertainties attending the indexing of complex powder patterns. It has been apparent that single-crystal data would be needed for a definitive structural determination.

In past preparation of sigma in the Co-Cr system by Mr. Nisbet of our laboratory, we have noted that the transformation to sigma was apparently complete and that the material was rather coarse-grained. Accordingly, attempts were made by Mr. Nisbet and Mr. Hadley to prepare single crystals of Co-Cr sigma. Such attempts have resulted in material which appears to be entirely sigma and in which we have found single crystals of irregular shape, generally about a few tenths millimeter in cross section.

Laue, oscillation, and Weissenberg photographs showed the crystals to be tetragonal with Laue symmetry $4/mmm$. The only systematic absences found were for Okl when $k+l=2n+1$, from which the possible space groups are $P4/mnm$, $P4nm$, or $P\bar{4}n2$.

The dimensions of the unit cell, determined from spectrometer traces of a powdered specimen are $a_0=8.75$A, $c_0=4.54$A. Analysis of the material gave 52.8 percent (atomic) Cr, and the density was determined as 7.98 ± 0.01 g/cc. The latter information gives 16 Cr and 14 Co atoms in the unit cell (calculated values—16.0 Cr, 14.3 Co).

Quite recently, we have learned of the work of Dickins, Douglas, and Taylor,[1] to be published shortly. They give essentially the same information as above for Co-Cr sigma. Furthermore, they point out the strong similarity of the sigma structural information to that of β-uranium,[2] reported by Tucker. We too find this similarity between β-uranium data and our results for Co-Cr.

On this basis we have tested a preliminary structure, similar to that proposed by Tucker for β-uranium. Since the latter is still in preliminary form and since our data are not sufficient to distinguish among the three possible space groups, we have chosen to consider only the centro-symmetric space-group $P4/mnm$ instead of $P4nm$ as used by Tucker. This implies flatness of hexagonal layers of atoms instead of the puckering proposed for β-uranium. Placing the atoms in the following positions of space-group $P4/mnm$,

2 in (b)
4 in (f), $x=0.100$
8 in (i), $x=0.373$, $y=0.027$
8 in (i), $x=0.573$, $y=0.227$
8 in (j), $x=0.300$, $z=\frac{1}{4}$,

we have made calculations of intensities for all reflections up to the CrK_α limit. There is satisfactory agreement with the observed powder pattern obtained with CrK_α to indicate the essential correctness of the proposed structure although refinement of parameters is indicated. Comparison with single crystal data cannot be made completely at this time—not only because of the incomplete state of our crystal data, but also because of large absorption effects for several reflections. Nonetheless, the reasonableness of the proposed structure as a preliminary for refinement by Fourier techniques is substantiated by the agreement of intensities wherever adequate comparison can be made.

Although there remains to be done the refinement of structure, particularly as to the questions of whether the hexagonal layers are flat, and the discrimination between Co and Cr positions, there seems little doubt that the sigma-phase structure is essentially that for β-uranium.

It is planned to obtain complete intensity data for Fourier investigations from more symmetrically shaped crystals, for which the absorption corrections could be more readily applied.

[1] Dickins, Douglas, and Taylor, Cavendish Laboratory, Cambridge, England. Private communication (1950) to C. W. Tucker, Jr.
[2] C. W. Tucker, Jr., Knolls Atomic Power Laboratory, Schenectady, New York. Private communication (1950): "An approximate crystal structure for the β-phase of uranium".

Cathodoluminescence of Zinc Orthosilicate with Manganese Activator

S. Larach and R. E. Shrader
Radio Corporation of America, RCA Laboratories Division, Princeton, New Jersey
December 18, 1950

Brown,[1] working at voltages between 200 and 800 volts, reported that the luminance of willemite, rbhdl.-Zn_2SiO_4:Mn, increased as the square of the voltage. Nottingham,[2] Martin and Headrick,[3] and Strange and Henderson[4] have reported variations with powers of voltage between one and two, for voltages up to 10 kv.

This letter reports measurements of the light output of rbhdl.-Zn_2SiO_4:Mn(0.3), extending the range of accelerating potentials from 10 to 50 kv for several screen thicknesses.

All screens were prepared by settling the phosphor particles through an aqueous solution of potassium silicate onto 1 in. × 3 in. glass slides. To avoid screen charging effects, the screens were aluminized.[5]

Measurements were made using a special demountable cathode-ray tube, details of which will be discussed in a later paper. The beam was focused and scanned over the whole screen in a standard television raster; current load was one microampere per square centimeter.

Figure 1 shows the variation of light output (at constant current) as a function of accelerating potential for three different screen thicknesses. It is found that for screens of this type and within this range of voltages, equations of the form $L=KV^n$ do

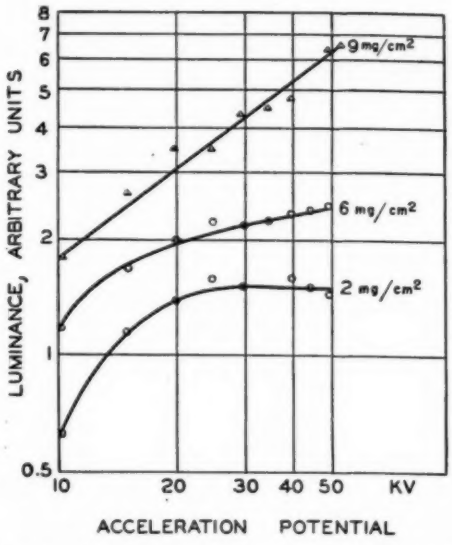


Fig. 1. Luminance as a function of accelerating potential for three thicknesses of rbhdl.-Zn_2SiO_4:Mn screens.

not fit the observed data, except possibly for the thickest screen (9 mg/cm²), where n is approximately 0.8.

Considering the appreciable efficiencies obtained for willemite (10 candle-power per watt), it would not be possible for an equation of the form $L=KV^n$ to apply for unlimited ranges of voltage for $n>1$, since this would eventually lead to a violation of the law of conservation of energy. The fact that $n>1$ occurs for the lower voltage ranges would indicate the interaction of complex internal phenomena.[6]

[1] T. B. Brown, J. Opt. Soc. Am. **27**, 186 (1937).
[2] W. B. Nottingham, J. Appl. Phys. **8**, 762 (1937).
[3] S. T. Martin and L. B. Headrick, J. Appl. Phys. **10**, 116 (1939).
[4] J. W. Strange and S. T. Henderson, Proc. Phys. Soc. (London) **58**, 383 (1946).
[5] D. W. Epstein and L. Pensak, RCA Rev. **7**, 1 (1946).
[6] H. W. Leverenz, *An Introduction to Luminescence of Solids* (John Wiley and Sons, New York, 1950), p. 443.

Investigations on the Capturing of Electrons in a 30-Mev Betatron

R. Wideröe
Brown, Boveri & Company, Baden, Switzerland
June 20, 1950

Recent publications[1] seem to indicate that the capturing process for electrons in the betatron is still open to discussion, there being no general accepted theory. It, therefore, seems worthwhile to mention a few experiments we have made with our 30-Mev-50-cycle betatron which strongly indicate that a special capturing theory—the space charge reduction theory—may be true.

At first we injected the electrons with an injection-voltage of relatively long duration as shown in Fig. 1. The parabolic curve 1 shows the "acceptance voltage", i.e., the voltage where the electrons will circulate on the middle equilibrium orbit in the vacuum tube; and the curves 2 and 3 are injection voltages used in our experiment. We found that the injection voltage 2 where $dU/dt>0$ at the moment when the injected voltage = the acceptance voltage (crossing A) gave no γ-ray yield at all. We had to move the injection voltage into the position of curve 3 so that $dU/dt<0$ (for $U_{inj}=U_{accep}$) to get any γ-ray yield. All our further experiments have established that the injection voltage must decrease at the moment when the electrons are to be captured.

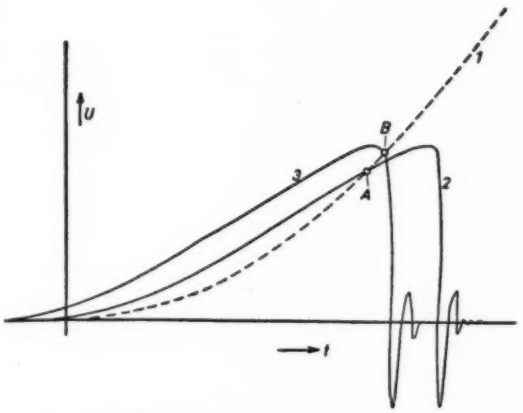


FIG. 1. Injection with an impulse voltage of long duration.

Our next investigation was to insert an electrode in a side arm of the vacuum chamber displaced approximately 180° from the electron gun. When electrons were injected into the vacuum chamber, it was found that a relatively strong current flows to this remote electrode that at no time could be reached by direct electrons from the gun. We, thus, have been led to the conclusion that the whole vacuum chamber must be filled up with electrons, probably produced by secondary, tertiary, etc., emission, the charge behaving like a very short-lived electron gas.

The electron gas builds up a space charge, and it is the reduction of this space charge when the injection voltage diminishes which helps to capture the electrons moving in the right equilibrium-orbits.

Kerst has suggested[2] that the magnetic field of the injected electrons may shrink the orbits and thus help to capture the electrons. Our experience shows that this effect (which should indicate electron capturing for an increasing injection voltage) must be smaller than the influence of the space charge from the electron gas.

Calculations also indicated that the shrinking of the electron orbit must be smaller than approximately 1 mm, but our experiments have shown that a plate extending 3 mm into the vacuum chamber did not reduce the γ-ray yield at all. I should mention that our target consisted of a small piece of platinum placed 15 mm inside the electron gun nearly half-way between the electron gun and the middle of the tube. It, thus, seems quite clear that in our experiments the reduction of space charge and not the shrinking effect of the magnetic field was responsible for the capturing of electrons.

However, we have also found a method to capture electrons by means of a different process, independent of the space charge.[3]

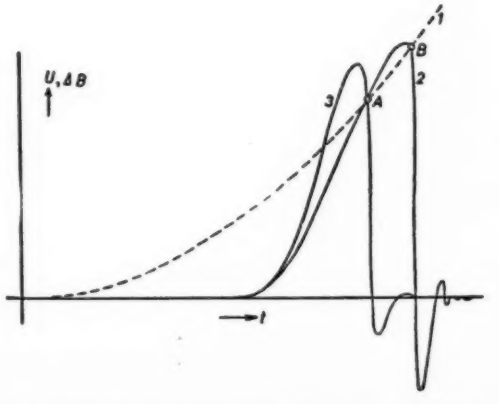


FIG. 2. Injection by means of a magnetic contraction field.

By superimposing over the guide field a rapidly changing magnetic field, the equilibrium orbit was contracted in about 0.2 μsec from its greatest value to the stationary orbit in the middle of the tube, and thus the electrons were trapped inside the electron gun and the target. By using a short injection voltage of only a few μsec duration (as shown in Fig. 2, curve 2), it was then possible to trap the electrons at the first crossing A with the acceptance voltage (curve 1) by means of the contraction field (curve 3) and also at the second crossing B by means of the diminishing electron gas. In this way it was possible nearly to double the γ-ray yield. The highest γ-ray intensity thus obtained was 150 r/min-m measured with a Victoreen ionization chamber behind a 3-mm lead screen. The injection voltage was approximately 44 kv_{max}, and the frequency 50 cycles per sec. The contraction field, which had to be precisely timed (to approximately ± 0.1 μsec) at the crossing point A, was energized by means of an impulse current flowing in a special coil surrounding only the guiding poles of the betatron.

[1] Jones, Kratz, Lawson, Ragan, and Voorhies, Phys. Rev. 78, 60 (1950). See also R. Wideröe, German Pat. Appl. No. W113742 VIIIc/21g-1/9/43, where a similar damping method has been proposed.
[2] D. W. Kerst, Phys. Rev. 74, 503 (1948).
[3] R. Wideröe, German Pat. Appl. No. W113563 VIIIc/21g-14/7/43.

The Thermal Expansion Coefficient and the Melting Point of Cubic Elements

D. VAN DER REYDEN
Statistical Section, Onderstepoort, South Africa
December 14, 1950

IN a recent issue of this Journal, M. E. Straumanis[1] demonstrated by means of a scatter diagram that the linear expansion coefficients of the cubic elements decrease with increasing melting points.

If it is assumed that the observed irregularities are due to inaccurate determination of the expansion coefficients at room temperature, and not due to some difference in the structure of the elements, this relationship should have some exact formulation. Within the range of the observed data, it should thus be possible to replace this, as yet unknown, law by some approximation.

Plotting the squared logarithms of the expansion coefficients against the squared logarithms of the melting points, a linear trend is obtained, thus yielding an empirical formula of the form: $y=\exp[a^2-b^2\log^2 x]^{\frac{1}{2}}$, where $y=\alpha\times 10^6$ and $x=T_F$ in °C.

Determining the numerical values of the parameters of this formula by the method of least squares $a=5.12$ and $b=0.6$, giving the smooth curve in Fig. 1.

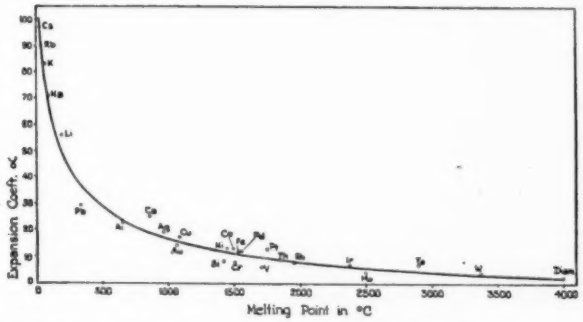


FIG. 1. Linear expansion coefficient $\alpha\times 10^6$ of cubic elements at room temperature *versus* melting point of the same elements.

This formula can be generalized to cover a wider range of data, if necessary. It would be interesting to derive the basic law and its approximation from first principles.

[1] M. E. Straumanis, J. Appl. Phys. 21, 936 (1950).

On the Semiconductor Bornite

C. W. HORTON
The University of Texas, Austin, Texas
December 12, 1950

SHERBINI AND YOUSEF[1,2] have made a few measurements on the dc current-voltage characteristics of bornite. Their data can be fitted with the formula[3]

$$I = A \exp(\alpha V) - 1, \qquad (1)$$

where $V = (V_{app} - Ir)$ of the simple diode theory. The data presented in the 1941 paper indicate the existence of both p-type and n-type conduction. These data can be fitted with the constants shown in Table I. The currents computed from Eq. (1) with the

TABLE I. Analysis of the data on bornite: Sherbini and Yousef, 1941.

T °C	α volts^{-1}	r ohms	A_p ma	A_n ma
30	6.3	60	0.288	0.0073
120	4.85	50	0.158	0.188
170	4.1	50	0.326	0.630

parameters of Table I differ from the 18 experimental values with an average error of 4 percent.

The data presented in the 1939 paper can be fitted by Eq. (1) with the parameters shown in Table II. Values of I computed from

TABLE II. Analysis of the data on bornite: Sherbini and Yousef, 1939.

T °C	α volts^{-1}	r ohms	A_p ma
−75	11.5	13.8	0.395
30	12.8	11.2	1.00

Eq. (1) agree with the 6 experimental values in the forward direction with errors of 0, 0, 2, 2, 25, and 34 percent. The two large errors occur for the smallest voltage in each case. The backward currents predicted by the equation have large percentage errors but small absolute errors. It appears that an n-type current exists for which A_n is of the order of 10^{-3} ma. The difference in r and A_p

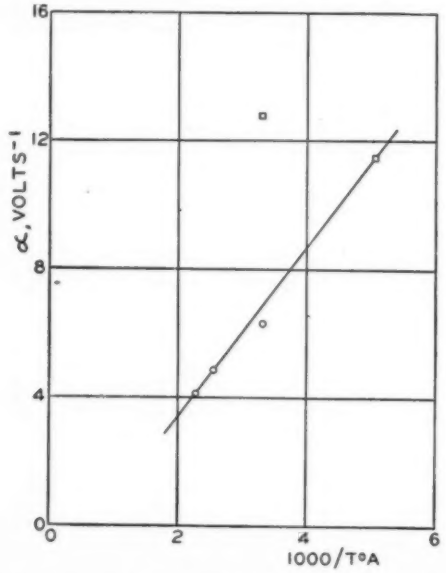


FIG. 1. The dependence of α on $1000/T°A$. ○ Sherbini and Yousef, 1941; □ Sherbini and Yousef, 1939.

for the two sets of values at 30°C is due, presumably, to a difference in the size of the copper whiskers.

Figure 1 is a plot of the five values of α *versus* $1000/T°A$. Four of these values fall on a straight line as they should, but there is no obvious reason for the erratic behavior of the fifth point.

Figure 2 is a plot of the values of A *versus* $1000/T°A$. The 1939

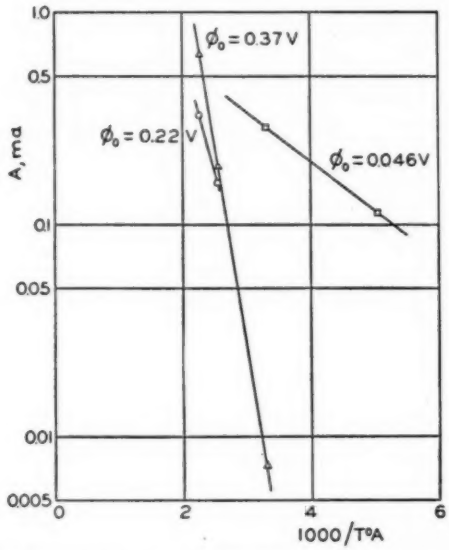


FIG. 2. The dependence of A on $1000/T°A$. ○ p-type, 1941; △ n-type, 1941; □ p-type, 1939. ϕ_0 is the deduced value of the contact potential difference.

data have been multiplied by 0.288, so that agreement is obtained at $T = 30°C$. The values of A_n fall on a straight line, while the values of A_p require two straight lines with different slopes. The corresponding values of ϕ_0 ($e\phi_0$ is the difference in the work functions) are indicated on the figure.

Since the work function of copper[4] is 4.9 electron volts, approximately, one arrives at the energy levels shown in Fig. 3 for bornite.

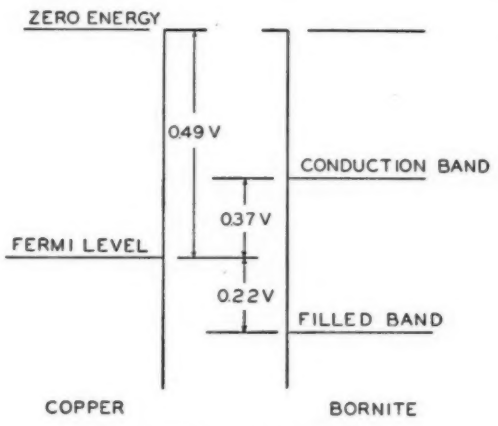


FIG. 3. The energy levels of bornite.

This interpretation is valid between 120°C and 170°C. It appears that the filled band is raised to a higher position below 30°C.

[1] M. A. El Sherbini and Y. L. Yousef, Proc. Phys. Soc. (London) **51**, 449 (1939).
[2] M. A. El Sherbini and Y. L. Yousef, Proc. Phys. Soc. (London) **53**, 120 (1941).
[3] Henry C. Torrey and Charles A. Whitmer, *Crystal Rectifiers* (McGraw-Hill Book Company, Inc., New York, 1948), p. 81.
[4] Newton Underwood, Phys. Rev. **47**, 502 (1935).

An Epitomization of the Basic Theory of the Generalized Schwarz-Christoffel Transformations as Used in Applied Physics

THOMAS JAMES HIGGINS
University of Wisconsin, Madison, Wisconsin
January 4, 1951

THE theory of functions of a complex variable encompasses certain branches of analysis which are among the most fruitful and powerful of the mathematical methods used in applied physics. Not the least of these methods is that of conformal mapping: namely, the representation of an area in the plane of a complex variable w upon an area in the plane of a second complex variable z in such fashion that the angle between two intersecting plane curves drawn in the w-area is the same as the angle between the representation of these curves in the z-area.

Perhaps the mapping problem most frequently encountered in applied physics and technology is to determine the function $w=f(z)$ which maps the area interior to, or exterior to, a given "polygon" in the w-plane on the upper half of the z-plane. Usually, the perimeter of the polygon is comprised of (1) line segments, some of which may be semi-infinite or infinite in length; (2) circular arcs; (3) a combination of line segments and one or more curvilinear arcs.

In that the problem of mapping the interior of a rectilinear polygon on the upper half of the z-plane has long been known as the Schwarz-Christoffel problem and the transformation $w=f(z)$ which solves it as the Schwarz-Christoffel transformation (after the two men[1,2] who independently solved the problem), it is convenient to encompass the stated group of mapping problems as the generalized Schwarz-Christoffel problem and the corresponding group of mapping functions as the generalized Schwarz-Christoffel transformations.

In view of the rapidly increasing use of generalized Schwarz-Christoffel transformations in high frequency electric circuit theory (three recent and particularly interesting applications occur in the work of Rice[3] on reflections from corners in rectangular wave guides, of Bucholz[4] on the impedance and attenuation of broadband cables comprised of two especially shaped "optimum" conductors, and of Trautman,[5] Huggins,[6] and others on the potential analog method of network synthesis) and of their long established usefulness in other branches of applied physics, an inclusive, integrated account of the theory of the generalized Schwarz-Christoffel transformations is most desirable. In, that no such account is available at present, however, those wishing to make such study must turn to assembling the desired information from the books and papers through which it is scattered. In this thought, the following epitomization of basic theory and account of the most useful sources of information is advanced as of interest and guidance to those who wish to study the theory in the most efficient manner.

1. *The rectilinear polygon.*—Four distinct cases occur:

i. The w-area to be mapped is interior to a polygon, and no corner point is to correspond to the infinite point on the real axis of z.
ii. The w-area to be mapped is interior to a polygon, but one corner point is to correspond to the infinite point on the real axis of z.
iii. The w-area to be mapped is exterior to a closed polygon, extends to infinity in all directions, and no corner point is to correspond to the infinite point of the z-plane.
iv. The w-area to be mapped is exterior to a closed polygon, extends to infinity in all directions, and one corner point is to correspond to the infinite point of the z-plane.

Case (i): Let $z_1, \cdots, z_n$ be the finite points on the real axis in the z-plane which correspond to the n corner points of the w-polygon, and let $\pi\theta_1, \cdots, \pi\theta_n$ be the internal angles at these corner points. Then the desired mapping function is

$$w=A\int\prod_{i=1}^{n}(z-z_i)^{1i-1}dz+B. \tag{1}$$

In this equation A and B are arbitrary complex constants, A fixing the scale and angular orientation and B the position of the polygon. Three corner points can be assigned specified positions on the real axis in the z-plane, providing only that these points occur in the same positive order on the real z-axis as do the corresponding corner points of the w-polygon as its perimeter is traversed in the positive direction (that is, the interior of the polygon is always to the left during traverse). When these three positions on the z-axis are chosen, the positions of the remaining corner points are fixed by the geometry of the w-polygon.

If $(\theta_i-1)>-1$, the corresponding corner point of the w-polygon is in the finite w-plane and the semicircle of infinite radius bounding the upper half of the z-plane corresponds to a point on one of the sides of the w-polygon. If $(\theta_i-1)=-1$, the corresponding "corner point" of the w-polygon comprises a rectilinear segment at infinity; the two sides of the w-polygon which "intersect" at the corner point are parallel semi-infinite lines; and the point z_i in the z-plane corresponds to the rectilinear segment at infinity, this being both perpendicular to and bounded by the two semi-infinite parallel lines. If $(\theta_i-1)<-1$, the corresponding "corner point" of the w-polygon comprises an arc at infinity; the two sides of the w-polygon which "intersect" at the corner point correspond to nonparallel semi-infinite lines; and the point z_i corresponds to the arc of the circle at infinity, this being both perpendicular to and bounded by the two semi-infinite nonparallel lines.

Case (ii): The desired mapping function is

$$w=A\int\prod_{i=1}^{n-1}(z-z_i)^{\theta i-1}dz+B, \tag{2}$$

where corner point w_n corresponds to the infinite point $z_n=\infty$ on the real axis of z, and two other z_i can be chosen as discussed above. If $i\neq n$, the corresponding z_i is finite and the corner point maps as discussed in case (i). For $(\theta_i-1)=-1$ or $(\theta_n-1)<-1$ the respective arcs of the circle of infinite radius on the w-plane correspond to the semi-infinite circle bounding the upper half of the z-plane. Contrariwise, if $(\theta_n-1)>-1$, the corner point is in the finite w-plane and corresponds to the infinite semicircle bounding the upper half of the z-plane.

Case (iii): The desired mapping function is

$$w=A\int\prod_{i=1}^{n}\frac{(z-z_i)^{\theta i-1}}{(z^2+1)^2}dz+B, \tag{3}$$

where now the $\pi\theta_i$ are the external angles of the w-polygon, hence the interior angles of the external area being mapped. The equations,

$$\sum_{i=1}^{n}\frac{z_i(\theta_i-1)}{z_i^2+1}=0 \quad \text{and} \quad \sum_{i=1}^{n}\frac{\theta_i-1}{z_i^2+1}=1, \tag{4}$$

must be satisfied in order that w be a single-valued function in the neighborhood of $z=i$ corresponding to the point at infinity in the w-plane. The phase of A and the value of B are determined by that point on the perimeter of the w-polygon (not a corner point) which is to be taken as corresponding to the point at infinity in the z-plane. The modulus of A and the n constants z_i are furnished by the set of equations

$$w_{i+1}-w_i=A\int_{z_i}^{z_{i+1}}\prod_{i=1}^{n}\frac{(z-z_i)^{\theta i-1}}{(z^2+1)^2}dz \quad (i=1, \cdots, n), \tag{5}$$

where w_i $(i=1, \cdots, n)$ are the corner points of the w-polygon.

Case (iv): The desired mapping function is

$$w=A\int\prod_{i=1}^{n-1}\frac{(z-z_i)^{\theta i-1}}{(z^2+1)^2}dz+B, \tag{6}$$

where now corner point w_n corresponds to the point at infinity in the z-plane.

The majority of technical books that discuss the Schwarz-Christoffel transformation for the rectilinear polygon give only the theory of case (i). Interestingly, they usually advance one or more illustrative examples which are of the type of case (ii); but no explanation is offered the reader why the term containing z_n is omitted. However, an excellent, complete treatment of both

cases (i) and (ii) is to be found in Guillemin's[7] recent book. A complete, but somewhat less detailed discussion, is given by Forsyth.[8] The numerous electrostatic problems solved by Thomson[9] serve to illustrate most of the points to be decided in setting up the transformation for a specific problem and determining the arbitrary constants A and B and the values of the z_i other than those which are assigned *a priori*. Cases (iii), (iv), or both—as noted in the references—are discussed by Bateman,[10] Jeffreys,[11] Villat,[12] Darboux,[13] and Harnack.[14] However, the papers of Kluyver[15] and Bickley,[16,17] who have developed the theory in detail, are most satisfactory for complete study of cases (iii) and (iv).

By virtue of the theory of analytic continuation, the function $w=f(z)$ can be continued across the real z-axis such that at conjugate points of the z-plane it provides conjugate values of w. Use of this property may considerably simplify determination of $w=f(z)$ for a polygon which has one or more axes of symmetry.

If the number of z_i is greater than the three which can be assigned arbitrarily as discussed above, the determination of specific values may prove a difficult task. It is to be noted that Schläfli[18] has advanced a systematic method of determining the values of the z_i.

The equations whereby the mapping is effected on the interior of a unit circle rather than on the upper half-plane are easily deduced through substitution for z from the well-known transformation mapping the upper half of the z-plane on a unit circle in, for example, the t-plane. In particular, Jeffreys[11] notes the advantage that may attend mapping of the outside of a polygon on the outside of a circle, Harnack[14] discusses mapping of the outside of a polygon on the inside of a unit circle, and Jeffreys[11] and Betz[19] consider mapping of the inside of a polygon on the inside of a unit circle.

2. *The circular-arc polygon.*—The problem of mapping the area interior to a w-polygon bounded by circular arcs on the upper half of the z-plane was solved by Schwarz.[1] The equations of the mapping function and the associated details are much more complicated than for the rectilinear polygon. In consequence, the reader is, perforce, referred to the excellent accounts of Forsyth,[8] Darboux,[13] and Betz.[19] The writer is not acquainted with explicit discussion of the mapping of the region external to the polygon. The complexities encountered in mapping circular-arc polygons and carrying through the details of transformation to where numerical calculation is possible are strikingly illustrated in papers by Daymond and Hodgkinson[20] and Fock.[21,22]

3. *Rectilinear-curvilinear polygons.*—If the curvilinear sides are circular arcs, the polygon can be considered as a subcase of Sec. 2, wherein the rectilinear sides correspond to arcs of infinite radius. In general, however, the theory pertinent to mapping a rectilinear polygon having one or more curvilinear sides is the theory of "curve-factors" developed by Leathem.[23–25] Account of this theory is to be found only in Leathem's original papers.

In conclusion, it is to be noted that Kober's[26] table of conformal transformations contains the mapping functions for numerous special rectilinear and circular-arc polygons. Betz's[19] book also contains a useful table of transformations.

[1] H. Schwarz, J. Math. (Berlin) **78**, 63–80 (1873).
[2] E. B. Christoffel, Ann. Math. (Rome) **1**, 95–128 (1867).
[3] S. O. Rice, Bell System Tech. J. **28**, 104–35 (1949).
[4] H. Bucholz, Arch. Elektrotech. **39**, 70–100 (1948).
[5] D. L. Trautman, Jr., *Amplifier Synthesis through Conformal Representation* (Ph.D. Thesis, Stanford University, 1949).
[6] W. H. Huggins, Proc. Inst. Radio Engrs. **36**, 421–4 (1948).
[7] E. A. Guillemin, *The Mathematics of Circuit Analysis* (John Wiley and Sons, New York, 1949), pp. 378–95.
[8] A. R. Forsyth, *Theory of Functions of a Complex Variable* (University Press, Cambridge, England, 1893), first edition, 537–54. Cases (i and ii).
[9] J. J. Thomson, *Recent Advances in Electricity and Magnetism* (University Press, Cambridge, England, 1893), chapter 3.
[10] H. Bateman, *The Partial Differential Equations of Mathematical Physics* (Dover Press, New York, 1942), pp. 294–309. Cases (i, ii, and iii).
[11] H. Jeffreys and B. Jeffreys, *Methods in Mathematical Physics* (University Press, Cambridge, England, 1946), pp. 395–8. Cases (i and iii).
[12] H. Villat, *Mecanique des Fluides* (Gauthier-Villars, Paris, 1938), chapter 3. Cases (i to iv).
[13] G. Darboux, *Leçons sur la Théories Générale des Surfaces* (Gauthier-Villars, Paris, France, 1887), Vol. 1, pp. 176–92. Case (i).
[14] A. Harnack, *Die Grundlaugen der Theorie des logarithmischen Potentiales* (G. Teubner, Leipzig, Germany, 1887), pp. 154–7.
[15] J. C. Kluyver, Z. Math. Phys. **40**, 129–50 (1895).
[16] W. G. Bickley, Trans. Roy. Soc. (London) **228A**, 235–78 (1928).
[17] W. G. Bickley, Proc. London Math. Soc. (2) **37**, 82–105 (1934).
[18] E. Schläfli, J. Math. **78**, 63–80 (1893).
[19] A. Betz, *Konforme Abbildung* (Springer-Verlag., Berlin, 1948), pp. 224–9, 251–9, 345–53.
[20] S. D. Daymond and J. Hodgkinson, Quart. J. Math. **10**, 136–44 (1938).
[21] V. Fock, J. Soc. Phys.-Math. (Leningrad) **1**, 147–68 (1926).
[22] V. Fock, Arch. Elektrotech. **16**, 331–40 (1926).
[23] J. G. Leathem, Trans. Roy. Soc. (London) **215A**, 438–87 (1915).
[24] J. G. Leathem, Proc. London Math. Soc. (2) **16**, 140–1 (1917).
[25] J. G. Leathem, Proc. Royal Irish Acad. **33**, 35–57 (1916–17).
[26] H. Kober, *Dictionary of Conformal Transformations* (Admiralty Computing Service, The Admiralty, London, 1948).